Kauno technologijos universitetas
Matematikos ir gamtos mokslų fakultetas

Vartotojų atsiliepimų internete veiksmingumą lemiančių veiksnių analizė

Baigiamasis magistro projektas

Joana Prasauskaitė
Projekto autorė

Doc. Dr. Vytautas Janilionis
Vadovas
Doc. Dr. Beata Šeinauskienė
Vadovė

Kaunas, 2018

Kauno technologijos universitetas
Matematikos ir gamtos mokslų fakultetas

Vartotojų atsiliepimų internete veiksmingumą lemiančių veiksnių analizė

Baigiamasis magistro projektas

Didžiųjų verslo duomenų analitika (621G12002)

Joana Prasauskaitė
Projekto autorė

Doc. Dr. Vytautas Janilionis
Vadovas
Doc. Dr. Beata Šeinauskienė
Vadovė

Doc. Dr. Tomas Ruzgas
Recenzentas
Dr. Asta Tarutė
Recenzentė

Kaunas, 2018

Kauno technologijos universitetas
Matematikos ir gamtos mokslų fakultetas
Joana Prasauskaitė

Vartotojų atsiliepimų internete veiksmingumą lemiančių veiksnių analizė

Akademinio sąžiningumo deklaracija

Patvirtinu, kad mano, Joanos Prasauskaitės, baigiamasis projektas tema „Vartotojų atsiliepimų internete veiksmingumą lemiančių veiksnių analizė“ yra parašytas visiškai savarankiškai ir visi pateikti duomenys ar tyrimų rezultatai yra teisingi ir gauti sąžiningai. Šiame darbe nei viena dalis nėra plagijuota nuo jokių spausdintinių ar internetinių šaltinių, visos kitų šaltinių tiesioginės ir netiesioginės citatos nurodytos literatūros nuorodose. Įstatymų nenumatytų piniginių sumų už šį darbą niekam nesu mokėjęs.

Aš suprantu, kad išaiškėjus nesąžiningumo faktui, man bus taikomos nuobaudos, remiantis Kauno technologijos universitete galiojančia tvarka.

(vardą ir pavardę įrašyti ranka) (parašas)

Turinys

Paveikslų sąrašas

Lentelių sąrašas

Prasauskaitė, Joana. Vartotojų atsiliepimų internete veiksmingumą lemiančių veiksnių analizė. Magistro baigiamasis projektas / vadovai doc. dr. Vytautas Janilionis, ir doc. dr. Beata Šeinauskienė; Kauno technologijos universitetas, Matematikos ir gamtos mokslų fakultetas.
Studijų kryptis ir sritis (studijų krypčių grupė): Taikomoji matematika (A02), Matematikos mokslai (A).
Reikšminiai žodžiai: interneto vartotojai, atsiliepimai, veiksmingumas, naudingumas, neigiama binominė regresija, klasifikavimas, dirbtiniai neuroniniai tinklai, artimiausių kaimynų metodas, atsitiktinių miškų metodas.
Kaunas, 2018. 47 p.

Santrauka

Vartotojų patirtys ir įžvalgos apie produktus ir paslaugas, pateiktos internetiniuose atsiliepimuose, padeda kitiems vartotojams apsispręsti dėl jų įsigijimo. Šiame darbe atsiliepimų veiksmingumas apibrėžiamas vartotojų atsiliepimų internete naudingumu. Naudojant neigiamą binominę regresiją nustatyta, kad atsiliepimai su neigiamu sentimentu mažina atsiliepimų naudingumą. Be to, restorano žvaigždutės įvertis, atsiliepimo dažnumas, ilgis ir atsiliepimų skaičius, teigiamai veikia atsiliepimų internete naudingumą. Stipriausiai atsiliepimų naudingumą veikia restorano žvaigždutės įvertis ir neigiamas atsiliepimo sentimentas. Atsiliepimai buvo klasifikuojami naudojant dirbtinių neuroninių tinklų, atsitiktinių miškų ir k-artimiausių kaimynų metodus. Geriausi rezultatai gauti, atsiliepimus klasifikuojant į naudingus ir nenaudingus, pritaikius dirbtinių neuroninių tinklų metodą

Prasauskaitė, Joana. The Analysis of the Factors Influencing the Performance of Online Consumer Reviews. Master‘s Final Degree Project / supervisors assoc. prof. Vytautas Janilionis and assoc. prof Beata Šeinauskienė; Faculty of School of Mathematics and Natural Sciences, Kaunas University of Technology.
Study field and area (study field group): Applied Mathematics (A02), Mathematical sciences (A)
Keywords: online consumers, reviews, performance, helpfulness, negative binomial regresion classification, Artificial Neural Networks, k-Nearest Neighbors, Random Forest.
Kaunas, 2018. 47 pages.

Summary

Consumers experience about goods and services submited to online consumer reviews, helps others consumers to make decision of buying them. In this research reviews performance of online reviews in terms of helpfulness. Using negative binomial regresion found that reviews, with negative sentiment receive less helpfulness. Moreover, the restauran star raiting, review length, frequency and number of a review, positively influence rewiew helpfulness. The most infulencing predictors of review helpfulness are restauran star raiting and negative sentiment of review. Also, reviews was classified in to helpful and not helpful. Classification experimented with artificial neural networks, random forests, k-nearest neighbors models and found that artificial neural networks produced better results.

Įvadas

Problematika. Vartotojų atsiliepimų internete veiksmingumo vertinimas ir klasifikavimas.

Tyrimo aktualumas. Internetinės vartotojų apžvalgos padeda vartotojams sužinoti teigiamus ir neigiamus aspektus apie produktą ar paslaugą, bei rasti tuos, kurie labiausiai atitinka jų poreikius. Tačiau verslui kyla sunkus uždavinys išanalizuoti atsiliepimus dėl jų charakteristikų, tokių kaip apimtis, įvairovė, greitis, bei teisingumas [1]. Visos minėtos charakteristikos būdingos didiesiems duomenims. Atsiliepimų internete analizės rezultatai padeda įmonėms pažinti vartotoją, įvardinti produktų ar paslaugų stiprybes ir silpnybes. Atsiliepimų veiksmingumo vertinimas yra žingsnis, tikslių klientų poreikių pažinimo link. Kita vertus, tai nauda ir patiems vartotojams, nes įmonės suinteresuotos ieškoti, bei taikyti naujus duomenų apdorojimo sprendimus, kurie klientams palengvintų rasti naudingą informaciją apie jų paslaugas, bei produktus. Atsiliepimų veiksmingumo vertinimo tyrimų rezultatai, naudojami kuriant vartotojų atsiliepimų internetinėse svetainėse rūšiavimo ir klasifikavimo automatines sistemas. Tai suteikia vartotojams patogesnę ir greitesnę paiešką, tarp didelio kiekio internetinėje svetainėje esamų atsiliepimų.

Tyrimo tikslas – nustatyti veiksnius, kurie lemia vartotojų atsiliepimų internete veiksmingumą ir juos panaudojus sukurti modelį vartotojų atsiliepimų klasifikavimui.

Uždaviniai:

1. Atlikti literatūros analizę ir atrinkti tyrime naudojamus kintamuosius, kurie gali būti reikšmingi atsiliepimų veiksmingumo tyrimui.
2. Pritaikius teksto tyrybos metodus paruošti atsiliepimų tekstus tolesnei analizei.
3. Nustatyti atsiliepimų sentimentus.
4. Regresinės analizės metodais nustatyti atsiliepimų internete veiksmingumą lemiančius veiksnius.
5. Sukurti atsiliepimų internete klasifikavimo modelius grįstus regresinės analizės rezultatais.

1. Literatūros apžvalga

Šiame skyriuje argumentuojamas vartotojų atsiliepimų internete veiksmingumą lemiančių veiksnių analizės aktualumas ir tyrimų problematika, atskleidžiama vartotojų atsiliepimų internete veiksmingumo samprata, identifikuojami veiksniai lemiantys vartotojų atsiliepimų internete veiksmingumą. Tai pat apžvelgiami matematiniai metodai, kurie tyrimuose buvo taikyti identifikuotai tyrimo problemai spręsti.

1.1. Vartotojų atsiliepimų internete veiksmingumą lemiančių veiksnių nustatymo aktualumas ir tyrimų problematika

Internetas tapo populiariausia ir lengviausiai prieinama terpe žmonėms reikšti savo nuomonę įvairiomis temomis. Informacija paremta paties vartotojo patirtimi ir įžvalgomis pateikiama internetinių atsiliepimų forma. Šių žinių milžiniškas srautas, kas dieną suplaukia į įvairius forumus ir svetaines. Dėl didelės pasiūlos rinkose dažnai sunku vartotojams apsispręsti ką verta rinktis, todėl ieškoma kitų vartotojų atsiliepimų internetinėje erdvėje, apie produkto ar paslaugos teigiamas ar neigiamas naudojimosi patirtis, savybes, kas padėtų išsirinkti labiausiai lūkesčius tenkinantį variantą. Taigi, naudingos kitų vartotojų pateiktos įžvalgos, patarimai, padeda klientams sumažinti riziką ir neapibrėžtumą susijusius su apsipirkimu internete [2]. Tarptautinės informacijos rinkimo ir matavimų bendrovės „AC Nielsen“, atlikto 50-tyje šalių tyrimo rezultatai parodė, kad 9 iš 10 vartotojų (90 %) pasitiki patarimu gautu iš šeimos narių ir aplinkinių rato [3]. O kitas šaltinis kuriuo pasitikima yra interneto vartotojų atsiliepimai (70 %). Verslas suvokia, kad internetiniais atsiliepimais yra skleidžiama informacija, kuri gali prisidėti prie vartotojų pasitikėjimo prekių ar paslaugų pardavėjais kūrimo, bei elektroninės komercijos skatinimo [4]. Dėl šios priežasties internetiniai atsiliepimai gali būti įrankiu, kuriant naujus rinkodaros sprendimus. Tereikia tinkamai panaudoti atsiliepimų forma vartotojų pateikiamą informaciją.

Atsiliepimo forma internete išreikštos žmonių įžvalgos teikia informaciją vartotojams apie produktą bei jo pardavėją, todėl jie taip pat daro įtaką vartotojų apsisprendimui, o kartu ir prekių, bei paslaugų pardavimams [5]. Sunku įvertinti vartotojų vyraujančias nuomones, kadangi atsiliepimais elektroninėje erdvėje pateikiamos informacijos srautai kas dieną tik didėja, tad vartotojų atsiliepimų internete tyrimai gali padėti įmonėms geriau suprasti vartotojų reakciją į jų produktus.

Atsiliepimų internetinėse svetainėse apdorojimą apsunkina juos apibūdinančios charakteristikos: apimtis, įvairovė, greitis, bei teisingumas (angl. *volume, variety, velocity, veracity*) [1]. Šiais keturiais požymiais apibrėžiami ir didieji duomenys (angl. *big data*) [6]:

- **apimtis** – didiesiems duomenis būdingas labai didelis duomenų kiekis, lyginant su tradiciniais duomenų šaltiniais;
- **įvairovė** – duomenys būna įvairaus formato, nes jų šaltiniai gali būti tiek žmonės, tiek mašinos;
- **greitis** – duomenų kiekis vis didėja, todėl duomenų apdorojimo greitis turi taip pat didėti;
- **teisingumas** – duomenys gaunami iš įvairių šaltinių, todėl reikia užtikrinti, kad jie būtų kokybiški ir patikimi.

Didžiųjų duomenų analizė taikoma vis plačiau. Žinoma, kad analitikai dirba su įvairių sričių duomenimis, tai yra sveikatos, verslo, saugumo, prekybos, logistikos, bei kitais. Stebimi pardavimai ir analizuojamas klientų elgesys, reikiama informacija gali būti renkama per lojalumo korteles, interneto naršymo istorijas, vartotojų komentarus. Į didžiųjų duomenų analizę verslas investuoja, nes tai įrankis naudojamas atsirinkti verslo plėtros kryptį ar atskirti tikslines vartotojų grupes, kurias domintų įmonės siūlomas produktas. Žinojimas kuriems vartotojams parduodamas produktas gali būti patrauklus, verslui leidžia sumažinti kaštus, sutaupyti laiko, bei padidinti pardavimus. Duomenų analizės pritaikymas atsiliepimams internete, gali padėti tobulinant produkto ar prekės kokybę, bei aiškinantis labiausiai vartotojų poreikius atitinkančio produkto charakteristikas. Vartotojų atsiliepimų internete rūšiavimo ir klasifikavimo problema taip pat sprendžiama duomenų analizės metodais. Siekiama, kad vartotojui būtų sukurtos sąlygos efektyviai atsakymų į rūpimus klausimus paieškai internetinių atsiliepimų gausoje, taip vartotojui palengvinant ir pagreitinant apsisprendimo priėmimą, dėl prekės ar paslaugos įsigijimo.

Tiek įmonės, tiek jų klientai nori susistemintų duomenų, suteikiančių apie tam tikrą produktą ar paslaugą naudingą pagrindinę informaciją, kurios šaltinis yra vartotojų atsiliepimai. Rinkodaros sprendimams vartotojų atsiliepimų analizė taip pat yra svarbi, kadangi jie veikia produktų pardavimus, bei įmonių rezultatus [2]. Literatūroje gausu mokslinių straipsnių, kurių tyrimo objektas yra interneto vartotojų atsiliepimų veiksmingumas (angl. *online consumers reviews performance)*. Viena iš aktualiausių šios tematikos mokslinių problemų – identifikuoti vartotojų atsiliepimų internete veiksmingumą prognozuojančius veiksnius [1,4,7]. Tokio pobūdžio rezultatai yra aktualūs internetinėms svetainėms kuriant automatines atsiliepimų internete rūšiavimo, bei klasifikavimo sistemas [1,7]. Analizuojant vartotojų atsiliepimų veiksmingumą M. Salehan'as ir D. J. Kim'as [1] nagrinėjo vartotojų internete atsiliepimų naudingumą (angl. *helpfulness*) ir skaitomumą (angl. *readership*) lemiančius veiksnius, panaudojant sentimentų analizę. Gauta, kad atsiliepimo veiksmingumą jo sentimentas veikia neigiamai, išskyrus teigiamą atsiliepimo

pavadinimo ir neutralų teksto sentimentus, kurie teigiamai veikia veiksmingumą. Be to nustatyta, kad skaitytojų dėmesį labiau patraukia ilgesni vartotojų atsiliepimai. Taip pat pastebėta, kad skaitytojai įvertina kaip naudingus ir tokius atsiliepimus, kuriuose nėra reikšmingos informacijos apie produktą ar paslaugą, todėl naudingumas gali netiksliai atspindėti tikrąjį atsiliepimų naudingumą. Be viso to, straipsnio autoriai M. Salehan'as ir D. J. Kim'as mano, kad ateities tyrimuose verta paanalizuoti, kaip įvairūs apsipirkimo internete aspektai, tokie kaip produkto ir paslaugų kokybė, bei kaina veikia vartotojų atsiliepimų internete veiksmingumą [1].

J. Wu savo darbe pateikia gaires, kaip galima įvertinti skirtingus atsiliepimo veiksmingumą reprezentuojančius aspektus ir atsiliepimo veiksmingumą vertinti šių aspektų bendru rezultatu [7]. Autorius priduria, kad tai pirmasis toks tyrimas kuriame kartu analizuojami du skirtingi, veiksmingumą apibūdinantys matai, tai yra atsiliepimo populiarumas ir naudingumas. J. Wu mano, kad jo sukurta sistema, naudojant Amazon.com internetinio puslapio vartotojų atsiliepimų duomenis, suteikia gilesnes įžvalgas veiksmingumo vertinime, nei konkuruojantys modeliai. Be to modelis gali padėti įvairiems sprendimus priimantiems asmenims veiksmingai naudotis vartotojų internete atsiliepimais [7].

Naudingumas dažniausias mokslinių darbų objektas tiriant atsiliepimų internete veiksmingumą. Modelio realizavimas, kuris automatiškai įvertintų naudingumą išspręstų atsiliepimų klasifikavimo problemą [1,4,8]. Atsiliepimų internete naudingumą savo darbe prognozavo T. L. Ngo-Ye'nas kartu su A. P. Sinha'nu įvertindami ne tik pačių atsiliepimų teksto, bet ir jų autoriaus charakteristikų, tokių kaip reputacija, įsipareigojimai, dabartinė veikla, įtaką atsiliepimų veiksmingumui [4]. Rezultatai parodė, kad ne tik atsiliepimo charakteristikas, bet ir informaciją apibūdinančią asmenį kuris parašė atsiliepimą, verta įtraukti prognozuojant atsiliepimų naudingumą. Tolimesniuose darbuose tyrimo autoriai mano, kad verta įtraukti atsiliepimą pateikiančio asmens socialinio profilio internete charakteristikas, tokias kaip draugų skaičius ar narystės trukmė. Modelis pasitvirtino naudojant restoranų atsiliepimų iš *Yelp.com* duomenų rinkinį, bei atsiliepimų duomenis apie knygas iš Amazon.com puslapio. Todėl T. L. Ngo-Ye'nas ir A. P. Sinha'nas tiki modelio pritaikomumu atsiliepimų iš kitų sričių naudingumo prognozavimui, tokių kaip elektronika, viešbučiai [4].

Moksliniai darbai atskleidžia, kad atsiliepimų internete veiksmingumo analizė yra aktuali tema. Vartotojų atsiliepimų internete veiksmingumas yra naujas reiškinys, kurio mokslinį pažinimą išplėstų šį reiškinį prognozuojančių veiksnių analizė. Tyrėjai skatina tęsti šios temos tyrimus, dėmesį sutelkiant į veiksnių, nuo kurių priklauso atsiliepimų internete veiksmingumas tyrimą, tai pat veiksmingumo prognozavimo, bei atsiliepimų klasifikavimo modelių kūrimą.

1.2. Vartotojų atsiliepimų internete veiksmingumo samprata

Vartotojų atsiliepimų internete veiksmingumo moksliniai tyrimai sutelkti į šį reiškinį lemiančias priežastis. Moksliniuose darbuose siekiama nustatyti, kokią įtaką vartotojų internete išreikštos nuomonės daro kitų vartotojų elgesiui ir sprendimams [9,10], taip pat nustatyti kokiais veiksniais remiantis gali būti prognozuojamas veiksmingumas [1,4]. Priklausomai nuo tyrimo konteksto vartotojų atsiliepimų internete veiksmingumas konceptualizuojamas labai įvairiai. Mokslinės literatūros analizė leidžia teigti, kad tai daugiadimensinis reiškinys apimantis atsiliepimų naudingumą, populiarumą, skaitomumą, ketinimą pirkti. Matai kuriais mokslinėje literatūroje vertinamas veiksmingumas pateikiami 1 lentelėje.

1 lentelė. Atsiliepimų veiksmingumo matai literatūroje

Veiksmingumo matas	**Straipsnis**	**Duomenys**
Naudingumas (angl. *helpfulness*)	M. Salehan'nas ir D. J. Kim'as (2016) [1], [4], W. Jianan'as (2017) [7], Y. Zhang'as and D. Zhang'as (2014) [13], J.Singh'as (2016) [8]	*Amazon.com*
	D. S. Yin'as, D. Bind'as and H. Zhing'as (2014) [11]	*shopping.yahoo.com*
	S. Karimi ir W. Fang'as. (2017), [12]	*play.google.com*
	T. L. Ngo-Ye'nas ir A. P. Sinha'nas (2014) [4], W. Jianan'as (2017) [7]	*Yelp.com*
Populiarumas (angl. *popularity*)	W. Jianan'as (2017) [7]	*Amazon.com ir Yelp.com*
Skaitomumas (angl. *readership*)	M. Salehan ir D. J. Kim (2016) [1]	*Amazon.com*
Ketinimas pirkti (angl. *purchase intention*)	F. R. Jimenez'as and N. A. Mendoza (2013) [9]	*Sukurtos manipuliacijos*
	J,Tian'a, Y. Chen'as and L, Wang (2014) [10]	*cnnic.com.*
Pardavimų pajamos (angl. *sales revenue*)	T. Liu'as and others. (2016) [14]	*Yahoo.com*

Šaltinis: sudarytas autoriaus

Plačiai tyrimuose [1,4,7,8,9,11,12] naudojamas matas vartotojų atsiliepimų internete veiksmingumui išmatuoti yra **naudingumas** (angl. *helpfulness*). Ar internetinėje erdvėje išsakytos žmonių nuomonės prisideda prie produkto vertinimo ir vartotojų sprendimo, gali būti nustatoma vertinant naudingumą [12]. Pavyzdžiui, prekių ar paslaugų apžvalgų įvertinimui *Amazon.com*, *tripadvisor.com* ir *Yelp.com* svetainėse naudojama balsavimo funkcija, kuri leidžia skaitytojui išreikšti nuomonę ar komentare pateikta informacija buvo jam naudinga. Atiduotas balsas, manoma prisideda prie asmens, kuris atsiliepimą įvertino kaip naudingą, sprendimo priėmimo dėl prekės ar paslaugos įsigijimo. Naudingi atsiliepimai internetinėse svetainėse gali būti iškeliami aukščiau kitų, kad vartotojai galėtų lengviau surasti kitų vartotojų nuomone naudingą informaciją. Tačiau visada ir visur norima tobulinti procesus, taupyti laiką, bei sąnaudas, todėl yra kuriami automatiniai internetinių atsiliepimų vertinimo ir klasifikavimo metodai. Y. Zhang'as ir D. Zhang'as [13] savo

moksliniame darbe pateikė modelį, kuris prognozuoja kiekvieno atsiliepimo apie produktą naudingumą (angl. *helpfulness*). Be to, darbo autoriai sukūrė algoritmą, kuris automatiškai suskirsto internetinius atsiliepimus į penkias klases: ypač naudingus, labai naudingus, šiek tiek naudingus, nelabai naudingus ir visai nenaudingus.

Jau minėta, kad atsiliepimų internete **naudingumas** vertinamas balsais. Atsiliepimo naudingumo balsai parodo žmonių skaičių, kuriems atsiliepime pateikta informacija buvo vertinga. Dažnai tyrimuose naudojama santykinė naudingumo įvertis, kuris apskaičiuojamas dalinant visų atsiliepimo balsų skaičių, kuriais manoma, kad atsiliepimas buvo naudingas, iš visų atsiliepimo balsų skaičiaus [1,7,12].

J. Wu'as savo darbe [7] apjungė atsiliepimo **naudingumo** ir **populiarumo** matus, pastarasis įvertina atsiliepimo potencialą pritraukti vartotojų dėmesį. Dažniausiai mokslinėje literatūroje veiksmingumas vertinamas naudojant vieną matą, vis dėl to tyrimo autorius J. Wu'as mano, kad naudojantis jo siūloma metodika, galima išsamiau įvertinti vartotojų atsiliepimų internete veiksmingumą [7].

Kitas mokslinėje literatūroje minimas veiksmingumo matas yra atsiliepimų **skaitomumas** (angl. *readership*). M. Salehan'as ir D. J. Kim'as [1] analizavo vartotojų dėmesio atkreipimo į atsiliepimą galimybę. Kadangi vartotojų atsiliepimų internete skaitomumas nėra tiesiogiai išmatuojamas, tyrimo autoriai skaitomumą įvertino pasitelkdami kiekvienam produktui parašytų atsiliepimų skaičių. Nors atsiliepimo perskaitymas yra pirmasis žingsnis vertinant jų veiksmingumą, vis tik tyrimuose skaitomumo aspektas beveik neanalizuojamas [1].

Interneto klientų **ketinimas pirkti** (angl. *purchase intention*) taip pat mokslinėje literatūroje randamas kaip atsiliepimų internete veiksmingumo matas [9,10] Vartotojų atsiliepimai prisideda prie pasitikėjimo produktais, jų pardavėjais, internetinėmis parduotuvėmis kūrimo. Verslui vartotojų pasitikėjimas yra svarbus, kadangi jam augant didėja ir ketinimo pirkti tikimybė. F. R. Jimenez'as ir N. A. Mendoza [9] tyrė internetinių atsiliepimų įtaką vartotojų ketinimui pirkti. Bendrai rezultatai parodė, kad patikimesnės apžvalgos yra tos, kurios išsamiau apibūdina produktą, o tai lemia didesnius pirkimo ketinimus. Kitame iš tyrimų tikrinama, ar pasitikėjimas grįstas emocijomis ir komentaro autoriaus pažinimu, daro įtaka vartotojų ketinimui pirkti [10]. Analizė patvirtino ir ankstesniuose darbuose gautus rezultatus, kad abu veiksniai daro teigiamą įtaką vartotojų ketinimui pirkti.

P**ardavimų pajamas** (angl. *sales revenue*) galima įvardinti, kaip dar vieną netiesioginį atsiliepimų veiksmingumo matą. T. Liu'o, X. Ding'o ir Y. Chen'o [14] tyrime, kino teatro pajamų prognozavimui naudojant internetinių atsiliepimų duomenis iš "Yahoo" filmų tinklalapio nustatyta,

kad internetiniai atsiliepimai daro įtaką filmų pajamoms. Tiksliausiai pavyko prognozuoti pardavimų pajamas, kai naudojami tokie veiksniai, ketinimo pirkti matas, atsiliepimo sentimentas, kino teatrų skaičius kuriuose bus rodomas filmas, bei direktoriaus populiarumas [14]. Literatūroje yra ir daugiau mokslinių darbų analizuojančių vartotojų atsiliepimų poveikį pardavimų apimtims [1,5].

C. Forman‘o ir kitų autorių darbe buvo nustatyta, kad informacijos apibūdinančios atsiliepimo autoriaus tapatybę atskleidimas skatina produktų pardavimus [5]. Be to, atsiliepimo autoriaus geografinės vietovės atskleidimas sustiprina šį ryšį. Straipsnio autorių manymu, norint padidinti internetinių verslų pardavimus, verta skatinti apžvalgų autorius pateikti daugiau apie save atskleidžiančio turinio.

Tai tik keletas atsiliepimų veiksmingumą apibūdinančių matų. Kadangi vartotojų atsiliepimų internete veiksmingumas yra aktualus tiriamas objektas, todėl ateityje tikėtina jų analizės metodai tobulės ir suteiks dar didesnę naudą verslui, bei vartotojams. Vis dėl to pastebima, kad daugiausiai dėmesio skiriama atsiliepimų naudingumo tyrimams. Remiantis atsiliepimų naudingumo mato rezultatais, yra kuriamos automatinės sistemos, kuriomis atskiriami atsiliepimai turintys vertingos informacijos ir esminių įžvalgų neturintys atsiliepimai (nereikšmingas tekstas) [1,4,8,13]. Taigi šiame darbe analizuojamas vartotojų atsiliepimų internete veiksmingumas yra apibrėžiamas kaip atsiliepimų naudingumas. Atsiliepimų internete naudingumo matas aprašomas vartotojų balsų skaičiumi, kurių nuomone atsiliepime pateikta informacija yra naudinga. Kuo didesnis naudingumo rodiklis, tuo didesniam skaičiui vartotojų, atsiliepimo autoriaus išsakytos mintys ir pastebėjimai buvo vertingi.

1.3. Vartotojų atsiliepimų internete veiksmingumą lemiantys veiksniai

Praeitame poskyryje apžvelgta kaip mokslinėje literatūroje konceptualizuojamas ir matuojamas vartotojų atsiliepimų internete veiksmingumas. Mokslinės literatūros analizė leido identifikuoti vartotojų atsiliepimų internete veiksmingumą lemiančius veiksnius, kurių apžvalga pateikta 2 lentelėje.

M. Salehan’as ir D. J. Kim’as [1] tyrimo rezultatai parodė, kad **teigiamas atsiliepimo pavadinimo sentimentas** (angl. *positive sentiment*) yra svarbus skaitomumo prognozavimo veiksnys, kitaip tariant, pozityvius pavadinimus turintys atsiliepimai yra dažniau skaitomi vartotojų. Atsiliepimo pavadinimo teigiamas sentimentas ir atsiliepimai su **neutraliu tekstu** (angl. *neutral text*) didina skaitomumą. Nustatyta, kad naudingumą neigiamai veikia atsiliepimo teksto sentimentas [1].

Atsiliepimo **gyvavimo laikas** (angl. *longevity*) įvertinamas skaičiuojant dienas nuo tada, kada atsiliepimas buvo sukurtas iki esamos datos. Nustatyta, kad atsiliepimo gyvavimo laikas daro teigiamą įtaką vartotojų internete apžvalgų skaitomumui [1]. T. L. Ngo -Ye'nas ir A. P. Sinha'nas [4] naudingumui prognozuoti panaudojo atsiliepimo charakteristiką panašią į gyvavimo laiką, tai yra atsiliepimo **naujumą** (angl. *recency*). Naujumas išmatuojamas dienų skaičiumi, tai yra skirtumas tarp seniausio atsiliepimo datos ir datos kada atitinkamas atsiliepimas buvo parašytas. W. Jianan'as [7] tirdamas atsiliepimų gyvavimo laiko įtaką skirtingų produktų atsiliepimų populiarumui, nustatė statistiškai reikšmingą, tačiau vieno produkto atsiliepimų populiarumui neigiamą, o kito produkto atsiliepimų populiarumui teigiamą ryšį. Taip pat tyrimas parodė, kad atsiliepimo gyvavimo laikas jo naudingumą veikia neigiamai, o naudojant kitus atsiliepimų duomenis nerastas statistiškai reikšmingas ryšys. Taigi reikia daugiau tyrimų, kurie atskleistų atsiliepimų gyvavimo laiko poveikį vartotojų atsiliepimų internete veiksmingumui.

Moksliniuose darbuose teigiama, kad ilgesni atsiliepimai teigiamai veikia atsiliepimų naudingumą, populiarumą ir skaitomumą [1,2,7]. Atsiliepimo **ilgis** (angl. *review length*) gali būti išmatuojamas žodžių skaičiumi atsiliepimo tekste [1,2,7]. Galima manyti, kad ilgesni atsiliepimai užima daugiau vietos svetainės naršymo lange, todėl jie yra greičiau pastebimi, nei trumpesnio teksto atsiliepimai. Vis dėl to, M. Salehan'o ir D. J. Kim'o tyrimo rezultatai rodo, kad ilgas atsiliepimo pavadinimas mažina atsiliepimų skaitomumą [1].

T. L. Ngo-Ye'no ir A. P. Sinha'nas [4] vartotojų **atsiliepimų** sukurtą **piniginę vertę (**angl. *monetary value***)** apibrėžė, kaip visų atsiliepimo autoriaus atsiliepimų naudingų balsų vidurkį. **Atsiliepimo dažnumas** (angl. f*requency*) yra visų atsiliepimo autoriaus ankščiau parašytų atsiliepimų skaičius. Be to, atsiliepimo teksto **žodžiai** yra atrinktas žodžių rinkinys, taip pat hipotetizuoti, kaip naudingumą lemiantys veiksniai. Remiantis [4] straipsniu, statistiškai reikšmingas ryšys tarp žodžių ir naudingumo vienuose atsiliepimuose gali būti žymiai stipresnis nei kituose, o tai gali priklausyti nuo verslo srities, kurioje atsiliepimai analizuojami konteksto ir atsiliepimų pobūdžio. Nustatyta, kad *atsiliepimo vertės, dažnumo ir naujumo* (angl. *monetary value, frequency, recency*) sąveika, teigiamai veikia naudingumą. Tai reiškia, kad ne tik atsiliepimo charakteristikos gali būti svarbios veiksmingumo prognozavimui, bet ir atsiliepimo autoriaus aktyvumą internetiniame puslapyje apibūdinantys veiksniai, tokie kaip atsiliepimo vertė, dažnumas ir naujumas [4].

W. Jianan'as [7] tyrė veiksmingumą lemiančius veiksnius, kurie apima atsiliepimų, jų autorių ir internetinės svetainės charakteristikas. Viena iš atsiliepimą apibūdinančių charakteristikų yra teksto **valentingumas** (angl. *valence*), kurio savybė pritraukti prasminiu požiūriu su juo susijusius kitų vartotojų komentarus. Darbe **valentingumas** yra išmatuojamas reitingo įverčiu (įvertis žvaigždute)

(angl. *star rating*), kuriuo atsiliepimo autorius įvertina prekę ar paslaugą. Valentingumą W. Jianan'as [7] detalizuoja dviem būdais. Pirmasis, kai valentingumas aprašomas penkiais nominalios skalės kintamaisiais (pvz. 1 žvaigždutė, 2 žvaigždutė,..., 5 žvaigždutė), o antruoju variantu, nurodomas kaip intervalų skalės kintamasis, tuomet valentingumo reikšmė lygi esamai atsiliepimo reitingo reikšmei. Rezultatai parodė, kad valentingumas neigiamai veikia atsiliepimo populiarumą, o naudingumą teigiamai, tačiau *valentingumo* ir *produkto tipo* sąveika daro neigiamą įtaką atsiliepimo populiarumui, tuo tarpu naudingumui statistiškai reikšmingas ryšys nerastas.

Atsiliepimų autoriaus **patikimumas** (angl. *credibility*), apibūdina vartotojo suvokimą ar atsiliepimo autorius atsiliepimuose pateikia teisingą informacija [7]. Patikimumą W. Jianan'as [7] savo darbe apibrėžia, kaip dvinarį fiktyvų kintamąjį (sukurtas dirbtinai), kuriam priskiriama reikšmė 1, kai informacija atsiliepime yra paremta atsiliepimo autoriaus patirtimi, kitu atveju priskiriama 0 reikšmė. Gauti tyrimo rezultatai parodė, kad atsiliepimo populiarumui jo autoriaus patikimumas daro neigiamą įtaką, o naudingumui teigiamą ir neigiamą įtaką, skirtingiems produktų atsiliepimų duomenų rinkiniams. Manoma, kad geresnis patikimumo mato panaudojimas būtų atsiliepimų autorių tapatybės atskleidimui tirti [7]. H. Hong'as, D. Xu su kitais tyrėjais, atlikę meta analizę, tai pat rado, kad atskleidžiama informacija apie atsiliepimo autorių, (tokia kaip tikras vardas, nuotraukos, gyvenamosios vietos atskleidimas) teigiamai veikia atsiliepimo naudingumą [2].

J. P. Singh'as su kitais darbo autoriais produkto **reitingą žvaigždute** savo tyrme įvardina, kaip bendrą produkto įvertinimo rodiklį [8]. Kitaip tariant, rodiklis yra visų vartotojų produkto įvertinimų vidurkis. Tyrimu nustatyta, kad produkto reitingas yra vienas iš svarbiausių parametrų atsiliepimų internete naudingumo vertinimui.

2 lentelė. Vartotojų atsiliepimų internete veiksmingumą lemiantys veiksniai

Veiksmingumą prognozuojantys veiksniai	**Veiksmingumo matas**	**Tyrimo rezultatas**	**Tyrimų autoriai**
Atsiliepimo ilgis	Naudingumas	Teigiamas ryšys	M. Salehan'as ir D. J. Kim'as, (2016) [1], H. Hong'as, D. Xu and others. (2017) [2], W. Jianan'as (2017) [7]
	Populiarumas	Teigiamas ryšys	W. Jianan'as (2017) [7]
	Skaitomumas	Teigiamas ryšys	M. Salehan'as ir D. J. Kim (2016) [1]
Atsiliepimo pavadinimo ilgis	Skaitomumas	Neigiamas ryšys	
Atsiliepimo pavadinimo sentimentas × teigiamas pavadinimas	Skaitomumas	Teigiamas ryšys	
Atsiliepimo sentimentas × neutralus atsiliepimas	Naudingumas	Teigiamas ryšys	
Atsiliepimo gyvavimo laikas (atsiliepimonaujumas, gylis)	Naudingumas	Nustatyti neigiami ir teigiami ryšiai [7] (tačiau dažniausiai nustatytas teigiamas ryšys)	H. Hong'as , D. Xu and others (2017) [2],W. Jianan'as (2017) [7]

2 lentelės tęsinys. Vartotojų atsiliepimų internete veiksmingumą lemiantys veiksniai

Veiksmingumą prognozuojantys veiksniai	Veiksmingumo matas	Tyrimo rezultatas	Tyrimų autoriai
Atsiliepimo gyvavimo laikas (atsiliepimo naujumas, gylis)	Populiarumas	Statistiškai reikšmingas (nustatyti teigiami ir neigiami ryšiai, skirtingiems atsiliepimų duomenų rinkiniams)	W. Jianan'as (2017) [7]
Atsiliepimo teksto žodžių rinkinys	Naudingumas	Teigiamas ryšys	T. L. Ngo-Ye'nas ir A. P. Sinha'nas (2014) [4]
Atsiliepimo vertė × dažnumas×gyvavimo laikas		Teigiamas ryšys	
Atsiliepimo valentingumas (junglumas)	Naudingumas	Teigiamas ryšys	W. Jianan'as (2017) [7]
	Populiarumas	Neigiamas ryšys	
Atsiliepimo valentingumas × produkto tipas	Naudingumas	Nerasta statistiškai reikšmingo ryšio	
	Populiarumas	Neigiamas ryšys	
Atsiliepimo autoriaus patikimumas	Naudingumas	Statistiškai nereikšmingas ryšys arba teigiamas ryšys	
	Populiarumas	Neigiamas ryšys	
Informacijos apie autorių atskleidimas (tikras vardas, gyvenamoji vieta, fotografija ir kita)	Naudingumas	Teigiamas ryšys	H. Hong'as , D. Xu and others (2017) [2]
Produkto reitingas žvaigždute	Naudingumas	Teigiamas ryšys	J. P. Singh'as and others (2016) [8]

Šaltinis: sudarytas autoriaus

Pastebima, kad dažnai mokslinėje literatūroje analizuojamas atsiliepimo teksto ilgis, kuris tyrimais nustatyta su vartotojų atsiliepimų internete naudingumu yra siejamas teigiamu statistiškai reikšmingu ryšiu [1,2,7]. Atsiliepimų teksto sentimentas tai pat įvardijamas, kaip veiksmingumą lementis veiksnys, tačiau su šia tekstą apibūdinančia charakteristika, naudojant skirtingų šaltinių atsiliepimus, gali būti gaunami skirtingi rezultatai. Veiksmingumo tyrimuose analizuojamos charakteristikos susijusios su atsiliepimų tekstu, jų autoriumi [1,7], bei verslo objektu [7].

1.4. Vartotojų atsiliepimų internete veiksmingumą lemiančių veiksnių tyrimui taikomų matematinių metodų apžvalga

Moksliniai darbai, kurių objektas vartotojų atsiliepimų internete veiksmingumas, dažniausiai sutelkti į jį lemiančių veiksnių nustatymą. Šiame darbe internetinių atsiliepimų veiksmingumas išmatuojamas naudingumo įverčiu. Toliau apžvelgiami analizės metodai, kurie mokslinėje literatūroje naudojami vartotojų atsiliepimų internete naudingumui tirti.

M. Salehan'as ir D. J. Kimb'as savo darbe [1], taikydami neigiamą binominę regresiją su logaritmine transformacija (angl. negative *binomial regression with logit transformation*), tyrė ar yra statistiškai reikšmingas ryšys tarp atsiliepimų sentimentų ir naudingumo. Analizuotas

Amazon.com 20-ties skirtingų produktų atsiliepimų rinkinys. Į lygtį įtrauktas fiktyvus kintamasis, aprašantis atsiliepimus su neutraliu poliškumu, kad išmatuoti konteksto kintamojo (moderatoriaus) poliškumą tarp atsiliepimo sentimento ir naudingumo. Fiktyvus kintamasis (neutralus atsiliepimas), tyrime apibrėžiamas kaip subalansuotas teigiamo ir neigiamo atsiliepimo sentimento lygis, bei gali įgyti vieneto reikšmę, kai atsiliepimas yra neutralaus poliškumo ir nulis kitais atvejais. Kiti regresijos lygties kintamieji yra atsiliepmo sentimentas, ilgis ir gyvavimo laikas. Kadangi kintamojo atsiliepimo ilgis reikšmių standartinis nuokrypis didesnis už jų vidurkį, todėl dispersijai sumažinti šis kintamasis regresijos lygtyje logaritmuotas. Atsiliepimo gyvavimo laikui taip pat pritaikyta logaritminė transformacija. Sentimento nustatymui darbe panaudota nemokama SentiStrength teksto sentimento nustatymo sistema (*SentiStrength* prieiga per: http://sentistrength.wlv.ac.uk/), kuri atskiria neigiamas ir teigiamas emocijas išreikštas tekste, suteikdama rangą nuo 1 iki 5 teigiamiems ir -1 iki -5 neigiamiems atsiliepimų sentimentams.

M. Salehan‘as ir D. J. Kimb‘as tyrime poliškumas apskaičiuotas taip: Poliškumas = teigiamas sentimentas + neigiamas sentimentas. O bendras atsiliepimo sentimentas: Sentimentas = (Teigiamas sentimentas + neigiamas sentimentas) – 2 ir jis matuojamas rangu nuo 2 iki 8. Taigi tyrime sudaryta regresijos lygtis [1]:

$$\begin{aligned}\frac{Naudingų\_atsiliepimų\_sk.}{Viso\_atsiliepimų\_sk.}\% &= \beta_0 + \beta_1 \cdot Atsiliepimo\_sentimentas \\ &+\beta_2 \cdot Neutralus\_atsiliepimas + \beta_3 \cdot \log(Atsiliepimo\_ilgis) \\ &+ \beta_4 \cdot Atsiliepimo\_sentimentas \times Neutralus\_atsiliepimas \\ &+ \beta_5 \cdot \log(atsiliepimo\_gyvavimo\_laikas),\end{aligned} \quad (1)$$

Naudingumas M. Salehan‘o ir D. J. Kimb‘o darbe išreikštas naudingumo ir visų atsiliepimų santykiu [1]. Tyrimo rezultatai parodė, kad statistiškai reikšmingas teigiamas ryšys yra tarp naudingumo ir atsiliepimo *ilgio*, *gyvavimo laiko*, bei *atsiliepimo sentimento* ir *neutralaus atsiliepimo* veiksnių sąveikos [1]. Taip pat nustatytas neigiamas ryšys tarp naudingumo ir atsiliepimo sentimento. Straipsnio autoriai M. Salehan‘as ir D. J. Kimb‘as teigia, kad dėl skirtumų tarp paslaugų ir produktų atsiliepimų, tyrimo išvadų negalima taikyti vartotojų atsiliepimams apie paslaugas [1]. Pastebima, kad tyrime lieka neįvertintas galimas sarkazmas atsiliepimuose. Taip pat yra atsiliepimų kurie įvertinti naudingumo balais, tačiau neturi vertingos esminės informacijos apie produktą, todėl atsiliepimų naudingumas gali ne visai tiksliai atspindėti realų.

T. L. Ngo-Ye’nas ir A. P. Sinha’nas [4] savo tyrime prognozuodami atsiliepimų naudingumą geriausius rezultatus gavo apjungus žodžių vektoriaus (angl. *Vector of words* arba *Bag of words model, BOW*) modelį ir RFM (angl. *recency, frequency, monetary value*) marketingo analizę. Vektorinės erdvės modelio rezultatas, tai iš atsiliepimo teksto atrinktų žodžių svoriai, kurie laikomi

naudingumo veiksniais. Realizuojant metodą, kiekvienas teksto dokumentas atvaizduojamas kaip žodžių vektorius, kuriame yra reikšmės nusakančios kiek kartų atitinkamas žodis pasikartojo tekste. Pavyzdžiui, žodžių krepšelis $\{t_1, t_1, t_{,1}, t_2, t_3, t_3\}$ aprašomas kaip vektorius $x = (3,1,2)$, kurio reikšmės apibūdina krepšelyje esamų žodžių dažnį. Tyrime gauti žodžių svoriai, tai kiekviename dokumente (atsiliepime) esančių žodžių dažnumas, kuris normalizuotas pagal dokumento ilgį [4] :

$$f_{i,j} \cdot \log \frac{dokumentų\_sk.}{sk.\_dokumentų\_su\_žodžiu\_i}, \tag{2}$$

čia $f_{i,j}$ yra žodžio i dažnis dokumente j.

BOW metodo trūkumai, labai didelis žodžių skaičius modelyje, gali pareikalauti daug kompiuterio resursų ir laiko apmokant modelį, o dar svarbiau tai gali sąlygoti modelio persimokymą [4]. Sprendžiant klasifikavimo uždavinį, gali būti netinkamai nustatomos klasės, kadangi žodžių krepšelio metodas nenaudoja žodžių junginių, o tik atskirus žodžius. Pavyzdžiui „not good“ žodis būtų priskiriamas neigiamai žodžių klasei, tuo tarpu atskiras žodis „good“ teigiamos klasės narys. Žodžiai kurie kartojasi labai retai, nesuteikia mokymosi metodui pakankamai informacijos teisingam atsiliepimų klasifikavimui, todėl tokie žodžiai apmokymo duomenyse yra ignoruojami. BOW metodas gali būti naudojamas kartu su kitais metodais, dažniausiai mašininio mokymosi, kaip pavyzdys naivaus Bajeso (angl. *naive Bayes*), atraminių vektorių (angl. *Support Vector Machines*) metodais [15].

RFM analizė yra tiesioginės rinkodaros metodas T. L. Ngo-Ye'no ir A. P. Sinha'no [4] tyrime vertina vartotojo, tai yra atsiliepimo autoriaus atsiliepimų *gyvavimo laiką* internetiniame puslapyje, *dažnumą* ir *pinginę vertę*. Šie trys įverčiai apibūdina atsiliepimo autoriaus įsitraukimo į internetinės bendruomenės veiklą stiprumą. Pritaikius atraminių vektorių regresijos algoritmą (angl. *Support Vector Regression Algorithm, SVR*), bei panaudojus žodžių svorius (BOW modelis) ir RFM analizės veiksnius, kaip kintamuosius atsiliepimų naudingumui prognozuoti, sudarytas tiksliausias modelis lyginant su rezultatais, kai BOW ir RFM metodai taikomi atskirai. T. L. Ngo-Ye'nas ir A. P. Sinha'nas išvadose teigia, kad tiek atsiliepimo teksto tiek jo autoriaus charakteristikos (gyvavimo laikas, dažnumas ir vertė) reikšmingai veikia vartotojų atsiliepimų internete naudingumą [4].

H. Hong'as, D. Xu su kitais autoriais atliko meta analizę, remiantis atliktų tyrimų apie naudingumą veikiančius veiksnius rezultatais. Kadangi skirtinguose darbuose gautos skirtingos išvados (vienuose tyrimuose nustatytas teigiamas ryšys, o kituose tas pats veiksnys nustatyta, kad turi neigiamą ryšį su atsiliepimų naudingumu) apie atsiliepimų naudingumo ryšį su įvairiais veiksniais, todėl norima suprasti ir suderinti tyrimuose gautas skirtingas išvadas [2]. Nustatyta, kad tyrimuose skirtingas išvadas daugiausiai lemia parinktas naudingumo vertinimo matas, produkto tipas ir

atsiliepimų šaltinis. Meta analizės rezultatai patvirtina, kad atsiliepimo teksto gylis, atsiliepimo amžius, recenzento informacijos atskleidimas ir recenzentų patirtis, turi teigiamos įtakos atsiliepimo naudingumui. Tai pat nustatyta, kad atsiliepimo skaitomumas ir atsiliepimo reitingas neturi reikšmingos įtakos apžvalgos naudingumui.

W. Jianan'as [7] santykinio naudingumo (santykis visų naudingų atsiliepimų, iš visų atsiliepimų, parašytų atitinkamam produktui ar paslaugai) vertinimui panaudojo Tobit regresiją. Logistinė regresija pritaikyta, kai priklausomas kintamasis apibrėžiamas atsiliepimo naudingų balsų skaičiumi. Tai pat logistinę regresiją S. Lee ir J. Y. Choeh'as naudojo atsiliepimų naudingumą veikiančių veiksnių tyrimui [16]. Atsiliepimai klasifikuoti naudojant sprendimų medžio, bei artimiausių kaimynų metodus. Geresnis klasifikatorius šiame tyrime buvo sprendimų medžio metodas [16].

Atsiliepimų naudingumo klasifikavimui J. P. Singh'as su kitais tyrimo autoriais pritaikė gradientinį stiprinimo algoritmą (angl. *gradient boosting algorithm*) [8]. Metodo mokymosi principas padalinti duomenis į mažas duomenų imtis. Tada modelis mokosi klasifikuoti duomenis kiekvienoje imtyje atskirai. Galiausiai rezultatai sujungiami ir gaunamas vienas klasifikavimo rezultatas. Lyginant su tiesine regresija, gradientinis stiprinimo algoritmas pateikia geresnius rezultatus [8].

R. Dong'as, M. Schaal'as su kitais tyrimo [17] autoriais, atsiliepimų klasifikavimui į naudingus ir nenaudingus taikė Naivaus Bajeso (angl. *Naive Bayes*), JRip ir atsitiktinių miškų (angl. *Random Forest*) metodus. Duomenų imtyje, naudingų ir nenaudingų atsiliepimų klases sudarė toks pats skaičius atsiliepimų. Geriausi klasifikavimo rezultatai gauti atsiliepimus klasifikavus atsitiktinių miškų metodu.

Y. Zhang'as, ir D. Zhang'as klasifikavo penkias atsiliepimų klases, kai naudingumas yra santykis tarp naudingų atsiliepimo balsų skaičiaus ir visų atitinkamo produkto atsiliepimų skaičiaus [13]. Kai atsiliepimo naudingumas didesnis nei 80% atsiliepimai priskiriami ypač naudingų atsiliepimų klasei, nuo 60% iki 80% atsiliepimai laikomi labai naudingais, 40% ir 60% intervale, kai kurie atsiliepimai yra naudingi. Naudingumo santykiui esant nuo 20% iki 40% atsiliepimai yra nenaudingi, o su mažesniu nei 20% naudingumo santykiu atsiliepimai priskiriami visai nenaudingų atsiliepimų klasei. Tyrime atsiliepimai klasifikuojami atraminių vektorių metodu kuris suklasifikuoja atsiliepimus 68.73% tikslumu [13].

P. J. Lee, Y. H. Hu ir K.T. Lu naudingus ir nenaudingus atsiliepimus klasifikavo sprendimų medžių, atsitiktinių miškų, logistinės regresijos ir atraminių vektorių metodais [18]. Tiksliausiai atsiliepimai klasifikuoti atsitiktinių miškų metodu.

Atsiliepimo naudingumo veiksnių įtakai nustatyti tyrimuose dažnai naudojami regresinės analizės metodai, bei jų modifikacijos. Atsiliepimų internete klasifikavimo uždaviniui spręsti naudojami mašininio mokymosi metodai, tokie kaip atsitiktiniai miškai, atraminiai vektoriai, sprendimų medžiai ir kiti. Pastebima, kad geru atsiliepimų klasifikavimu pasižymi atsitiktinių miškų metodas [17,18].

1.5. Programinių kalbų apžvalga

R yra atviro kodo programa, skirta statistiniams skaičiavimams, bei ekonometriniam modeliavimui. R programavimo kalbą 1995 m. sukūrė R. Ihaka ir R. Gentleman'as. Dirbant R, galima naudotis plačiu statistinių (tiesinio ir netiesinio modeliavimo, klasikinės statistinės analizės, laiko eilučių analizės, klasifikavimo, klasterizavimo ir kt.), teksto tyrybos ir grafinių metodų įvairove. Programavimo kalbos aplinka sukurta R, Fortan ir C kalbomis. Be to, R turi panašumų su kitomis kalbomis, tokiomis kaip Java, C, ir Perl, o tai suteikia prieigą naudotojams prie įvairių komandų atliekant skaičiavimo ir analizės užduotis. R kalba yra paprasta, todėl vartotojui greitai perprantama. Šiai programinei įrangai skirti metodai gali būti vartotojų papildomi, kuriami ir tobulinami. Skelbiama, kad CRAN repozitorijoje yra sukurta daugiau nei 10 tūkstančių programinių paketų, kurių skaičius su kiekvienais metais eksponentiškai auga [21]. 2017 metais R kalba programavimo kalbų reitingo lentelėje užėmė 14 vietą [19].

Kita programavimo kalba, kuri naudojama ne tik duomenų analizės, bet ir kitų sričių specialistų yra Python. Python yra labai sparčiai populiarėjanti kalba, programavimo kalbų reitinge 2017 metais ji užėmė trečią vietą [19]. Didžiausias Python kalbos privalumas yra didžiulė standartinių funkcijų biblioteka, kuri leidžia pagreitinti, bei supaprastinti programų kūrimą, tai pat kurti universalias ar pagal poreikius pritaikytas programas. Ši programavimo kalba yra paprasta, norimą atlikti veiksmą galima užrašyti keliais variantais. Vienas iš Phyton trūkumų yra lėtas programos vykdymo greitis.

SAS yra specializuota programavimo kalba, skirta statistinei duomenų analizei. Skirtingai nuo integruotų įrankių, prieinamų iš tokių programų kaip „Microsoft Excel", SAS leidžia vartotojams gauti ir tvarkyti duomenis iš įvairių šaltinių ir suteikia daug daugiau kontrolės ir laisvės kompiliuojant duomenis. Ši programinė įranga leidžia automatiškai kurti statistinės analizės rezultatų ataskaitas ir jas išsaugoti HTML, PDF ir RTF formatais. SAS programinė įrangą naudojama versle, sveikatos tyrimuose, gamyboje.

Matlab programinė įranga skirta įvairių mokslo sričių problemoms spęsti, tačiau daugiausiai matematinėms. Pirminis tikslas, kuriant šia programinę įrangą buvo sukurti įrankį darbui su matricomis. Tačiau dabar Matlab yra galingas paketas, turintis savitą, tačiau lengvai perprantamą programavimo kalbą. Tai kalba apimanti iš anksto paruoštas matematines funkcijas. Matlab

programinis paketas naudojamas matematiniams skaičiavimams, duomenų surinkimui, imitaciniam modeliavimui ir maketavimui, algoritmų sudarymui, mokslinei ir inžinerinei grafikai, duomenų analizei, bei jos rezultatų vizualizacijai.

IBM SPSS Statistics yra specializuota statistinė programinė įranga, skirta statistinei duomenų analizei. Ji leidžia atlikti visą duomenų analizės procesą, tai yra įkelti duomenis iš įvairių šaltinių, juos paruošti (atlikti transformacijas, sukurti naujus kintamuosius, užkoduoti kategorijas ir kita), išanalizuoti duomenis tiksliais statistiniais metodais, lengvai ir suprantamai pateikti gautus rezultatus grafikais, bei analitinėmis lentelėmis, eksportuoti rezultatus įvairias formatais. SPSS skirta statistinei analizei yra pritaikyta atlikti teksto analizę, be to ji turi mašininio mokymosi metodų biblioteką.

1.6. Tyrimo tikslas ir uždaviniai

Tyrimo tikslas – nustatyti veiksnius, kurie lemia vartotojų atsiliepimų internete veiksmingumą ir juos panaudojus sukurti modelį vartotojų atsiliepimų klasifikavimui.

Uždaviniai:

1. Atlikti literatūros analizę ir atrinkti tyrime naudojamus kintamuosius, kurie gali būti reikšmingi atsiliepimų veiksmingumo tyrimui.
2. Pritaikius teksto tyrybos metodus paruošti atsiliepimų tekstus tolesnei analizei.
3. Nustatyti atsiliepimų sentimentus.
4. Regresinės analizės metodais nustatyti atsiliepimų internete veiksmingumą lemiančius veiksnius.
5. Sukurti atsiliepimų internete klasifikavimo modelius grįstus regresinės analizės rezultatais.

2. Atsiliepimų klasifikavimas grįstas regresinės analizės rezultatais

Skyriuje pateikiama tyrime naudojama medžiaga ir metodų aprašymai.

2.1. Kintamųjų apibrėžimai

Tyrime naudojami duomenys yra iš *Yelp.com* svetainės. Atsižvelgus į turimus duomenis, bei literatūros analizę, darbe pasirinkti tirti tokie veiksniai, kurie gali daryti įtaką vartotojų atsiliepimų naudingumui:

1. Atsiliepimo **valentingumas** – matuojamas atsiliepimo žvaigždučių sistemos įvertinimu. (angl. *review star*).
2. Atsiliepimų **skaičius** – kiekvienam restoranui svetainėje parašytų atsiliepimų skaičius (angl. *review count*)
3. Atsiliepimo **naujumas** – skirtumas dienomis, tarp seniausio atsiliepimo datos ir i-tojo atsiliepimo datos (angl. *review recency*).
4. Atsiliepimo **dažnumas** (angl. *review* f*requency*) – atsiliepimų skaičius, kuriuos vartotojas buvo parašęs prieš i-tąjį atsiliepimą.
5. **Restorano vertinimas žvaigždute** (angl. *star rating*) – šiuo įverčiu įvertinama bendra vartotojų nuomonė apie restoraną, kadangi tai yra visų restorano internetinėje svetainėje vertintojų, įvertinimų vidurkis.
6. Atsiliepimo teksto **ilgis** (angl. *review length*) – atsiliepimo tekste esamų žodžių skaičius.
7. Atsiliepimo **sentimentas** (angl. *review sentiment*) – atsiliepimų teksto neigiamas, neutralus ir teigiamas sentimentas.

Iš turimų duomenų, sukuriami nauji kintamieji *naujumas, dažnumas, ilgis,* ir *sentimentas*. Toliau pateikiama, kaip minėti kintamieji apskaičiuoti, bei kokios tam naudotos R programinio paketo funkcijos.

Atsiliepimo naujumas. Skaičiuojant atsiliepimo naujumą, panaudota **as.Date** funkcija, kuri datos reikšmes paverčia į datos formatą. Tada apskaičiavus skirtumą tarp naujausio atsiliepimo datos ir i -tojo atsiliepimo gaunamos atsiliepimo naujumo reikšmės.

Atsiliepimo dažnumas. Pirmiausia, R programoje duomenys surūšiuojami naudojant ***order*** funkciją. Naudojant ***order*** funkciją rūšiuojamos duomenų eilutės pagal *user_id,* po to pagal *Data* kintamąjį, mažėjimo tvarka. Visų atsiliepimo autorių atsiliepimai surūšiuoti nuo seniausio iki naujausio. Kitu žingsniu, pradedant vienetu, sunumeruotas kiekvienas unikalaus *user_id* atsiliepimas (eilutės), tam panaudojus ***sequence*** numeravimo funkciją. Gauta kiekvieno atsiliepimo charakteristika parodo, kiek atsiliepimų autorius buvo parašęs prieš i-tąjį atsiliepimą.

Atsiliepimo ilgis. Su ***unlis*t** ir ***strsplit*** funkcijomis atsiliepimo tekstas išskaidomas į atskirų žodžių vektorių, ***lenght*** *ir* ***sapply*** funkcijų pagalba suskaičiuojamas kiekvieno atsiliepimo teksto žodžių skaičius.

Teksto sentimentų nustatymo žingsniai pateikti sekančiame poskyryje.

2.2. Teksto apdorojimas ir sentimentų analizė

Atsiliepimų naudingumas, tai pat gali priklausyti nuo atsiliepimo teksto sentimento. Todėl atsiliepimų sentimentas yra dar vienas šiame darbe analizuojamas veiksnys. Sentimentų analizė dažniausiai atliekama su išvalytu tekstu, todėl atlikti tokie atsiliepimų teksto valymo žingsniai:

- pašalinami skyrybos simboliai, tam naudojama *removePunctuation* R paketo procedūra;
- pašalinami skaičiai, *removeNumbers* procedūra;
- tekstas paverčiamas į mažąsias raides, naudojant *tolower* procedūrą;
- pašalinti pridėtiniai tarpai tarp žodžių, *stripWhitespace* procedūra;
- pašalinami visi simboliai, kurie nėra anglų kalbos abėcėlėje;
- pašalinami nereikšmingi žodžiai, jungtukai dalelytės, kurie yra žodyne myStopwords (pavyzdžiui: „was“, „it“, „during“, „before“), papildomai į šį žodyną įtaukta „xl“, „two“, „just“, „las“, „vegas“, „le“ , „ive“;
- atliekamas teksto lematizavimas (*lemmatize_words* procedūra), kuris reiškia, kad skirtingų gramatinių formų žodžiai, bus verčiami į esamojo laiko žodžių formą (pavyzdžiui žodžiai „run“, „ran“ , „runing“ žodžiai paverčiami į „run“).
- išskiriamas žodžių kamienas naudojant *stemDocument* procedūrą.

Dažniausiai sentimentai skirstomi neigiamus ir teigiamus. Tai pat gali būti ir trečioji, neutralaus sentimento kategorija. Sentimentas nustatomas naudojant sudarytą žodžių žodyną, kuriame yra kiekvieną sentimento kategoriją apibūdinantys žodžiai. Žodyną galima susidaryti pačiam, tačiau yra daugybė jau sudarytų. Sentimentų ištraukimui naudojama R programinio paketo *analyzeSentiment* procedūra ir *QDAP* bendrinis žodynas. Neigiamą sentimentą turinčiam atsiliepimui priskiriamas žodis „negative“, neutraliam „neutral“, o teigiamam „positive“. Toliau darbe priskiriamos reikšmės -1 neigiamą, 0 neutralų ir 1 teigiamą sentimentą turinčiam atsiliepimui.

2.3. Neigiama binominė regresinė analizė

Naudojant regresinę analizę, galima įvertinti priklausomo kintamojo priklausomybę nuo nepriklausomų kintamųjų, bei prognozuoti priklausomojo kintamojo reikšmes. Šiame darbe nepriklausomas kintamas *Y* yra atsiliepimo naudingumas. Binominė regresija naudojama, kai priklausomas kintamasis nėra pasiskirstęs pagal normalųjį skirstinį. Taigi, modelis skirtas

modeliuoti dispersijos perviršį, kai priklausomo kintamojo Y dispersija yra daug didesnė už jo vidurkį. Be to, regresijos lygtis sudaroma ne pačiam priklausomam kintamajam Y, o jo vidurkiui μ [16]:

$$\begin{aligned} &\mu = e^{z}, \\ &z = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_n x_n, \end{aligned} \tag{3}$$

čia: μ – priklausomo kintamojo Y vidurkis, z – vidutinis įvykių skaičius, $e = 2.718...$, $x_1,...,x_n$ – nepriklausomi kintamieji, $\hat{\beta}_0, \hat{\beta}_1 ..., \hat{\beta}_n$ – modelio koeficientai kur $\hat{\beta}_0$ yra modelio konstanta.

Kai nepriklausomas kintamasis x_i padidėja vienetu, priklausomo kintamojo Y vidurkis μ padidėja (sumažėja) e^{β_i} karto [20]. Taigi parametro koeficientas β_i, parodo nepriklausomų kintamųjų teigiamą ar neigiamą poveikį Y vidurkiui.

Neigiamos binominės regresijos modelio etapai susideda iš kintamųjų tinkamumo tyrimo ir modelio tinkamumo duomenims tikrinimo [20]:

1. Patikrinama ar priklausomo kintamojo Y dispersija didesnė už vidurkį. Jei dispersija mažesnė, neigiama binominė regresija netaikoma.
2. Sudaromas neigiamos binominės regresijos modelis ir tikrinama, ar deviacijos ir jos laisvės laipsnių santykis artimas vienetui.
3. Patikrinama, ar tikėtinumų santykio kriterijaus $p > 0{,}05$. Jei yra kitaip, modelis netinkamas.
4. Patikrinama, ar visi nepriklausomi kintamieji yra statistiškai reikšmingi (visos Voldo kriterijaus, išskyrus konstantą reikšmės $p > 0{,}05$). Jei nors vienas nepriklausomas kintamasis nėra statistiškai reikšmingas modelis dar netinkamas naudoti prognozavimui.

Būtina patikrinti ar sudarytas regresijos modelis tinkamai aprašo priklausomybę tarp priklausomo kintamojo Y ir nepriklausomų kintamųjų $x_1,...,x_n$. Modelio tinkamumą duomenims aprašančios charakteristikos [20]:

- *Deviacijos* rodiklis parodo kiek tiriamas modelis skiriasi nuo modelio pilnai aprašančio duomenis (kuris netinkamas naudoti nes yra labai sudėtingas). Regresijos modelis laikomas pakankamai geru, kai deviacija padalinta iš savo laisvės laipsnių, nuo vieneto skiriasi nedaug. Priimtinas intervalas 0,9 – 1,1, jeigu reikšmė mažesnė už 0,7 ar didesnė už 1,3 tai parodo blogą modelio tikimą duomenims.

- *Pirsono chi kvadrato statistika* (angl. *Pearson Chi – Square*) yra deviacijos analogas, kuris matuoja ar su regresijos modeliu gautų ir tikrų Y reikšmių skirtumai yra maži, tai yra statistiškai nereikšmingi. Laikoma, kad modelis pakankamai geras, jeigu chi kvadrato statistikos ir savo laisvės laipsnių santykis, nedaug skiriasi nuo vieneto ir bent patenka į intervalą 0,8 –1,2.

- *Tikėtinumo santykio chi kvadrato kriterijus* lygina sudarytą modelį su modeliu neturinčiu jokių nepriklausomų kintamųjų ir tikrina ar modelyje yra bent vienas reikšmingas kintamasis (regresorius). Kai $p > 0{,}05$, tai regresijos modelis gali abejotinai aprašyti duomenis. Kai ši statistika didelė ji rodo geresnį modelio tinkamumą.

- *Akaikės informacinis kriterijus (AIC)* leidžia palyginti du modelius. Geresnis modelis laikomas tas, kurio AIC mažesnis. Lyginamųjų modelių parametrų įverčiai turi būti suskaičiuoti, naudojant tuos pačius duomenis. Taip pat vieno modelio nepriklausomi kintamieji turi būti kito modelio nepriklausomų kintamųjų aibės dalis. AIC informacinis indeksas apskaičiuojamas taip [21]:

$$AIC = \frac{2k}{n} + \log\left(\frac{RSS}{n-k}\right), \tag{4}$$

čia k – nepriklausomų kintamųjų skaičius,

n – stebėjimų skaičius,

$RSS = \sum_i e_i^2$ – liekamųjų paklaidų kvadratų suma, kur $e_i = y_i - \hat{y}_i$, kai y_i – i-toji priklausomojo kintamojo reikšmė ir $\hat{y}_i$ – i-toji priklausomojo kintamojo reikšmė, apskaičiuota pagal sudarytą modelį.

Kai į lygtį norima įtraukti kategorinius kintamuosius, sukuriami vadinamai pseudokintamieji, kurie gali į gyti tik dvi reikšmes 0 ir 1. Atsiliepimo sentimentas yra kintamasis turintis 3 kategorijas, todėl sukuriami du dvireikšmiai pseudokintamieji. Vienas jų *teigiamas sentimentas,* kuris įgyja reikšmę 1, kai sentimentas yra teigiamas ir 0 kitais atvejais, bei pseudokintamasis *neigiamas sentimentas* įgyjantis reikšmę 1, kai atsiliepimas yra neigiamo sentimento ir 0 kitais atvejais.

Neigiamai binominei regresijai darbe realizuoti, naudojama R programos ***glm.nb*** komanda, kuri yra iš **MASS** funkcijų bibliotekos. Didžiausio tikėtinumo chi kvadrato statistika tikrinama naudojant **anova** komandą.

2.4. Vartotojų internete atsiliepimų klasifikavimo metodai

Klasifikavimas (angl. *classification*) – tai objektų skirstymas į klases. Klasifikavimo uždavinio tikslas – turimos duomenų aibės klasifikavimui (klasės yra žinomos) sukurti taisyklę, pagal kurią duomenys, kurie nubuvo naudojami klasifikavimo taisyklės kūrime, automatiškai būtų priskiriami vienai ar kitai žinomai klasei.

2.4.1. Dirbtiniai neuroniniai tinklai

Dirbtiniai neuroniniai tinklai (angl. *Artificial neural networks, ANN),* yra vienas mašininio mokymosi metodų. ANN galima naudoti klasifikavimo uždaviniui spręsti ar realių reikšmių prognozavimui. Dirbtinis neuroninis tinklas – rinkinys tarpusavyje sujungtų neuronų. Neuroniniai tinklai dažniausiai sudaryti iš trijų sluoksnių, tai yra įvesties sluoksnio (angl. *input*), paslėpto neuronų sluoksnio (angl. *hidden*) ir išvesties sluoksnio (angl. *output*). Kiekvienas neurono sluoksnis yra sujungtas su gretimų sluoksnių neuronais, tačiau dažniausiai to paties sluoksnio neuronai nėra sujungti. Neuronų skaičiavimo rezultatas išvesties sluoksnyje priklauso nuo paslėptųjų sluoksnių skaičiaus, neuronų išvedimo reikšmių ir su jais susijusių svorinių koeficientų.

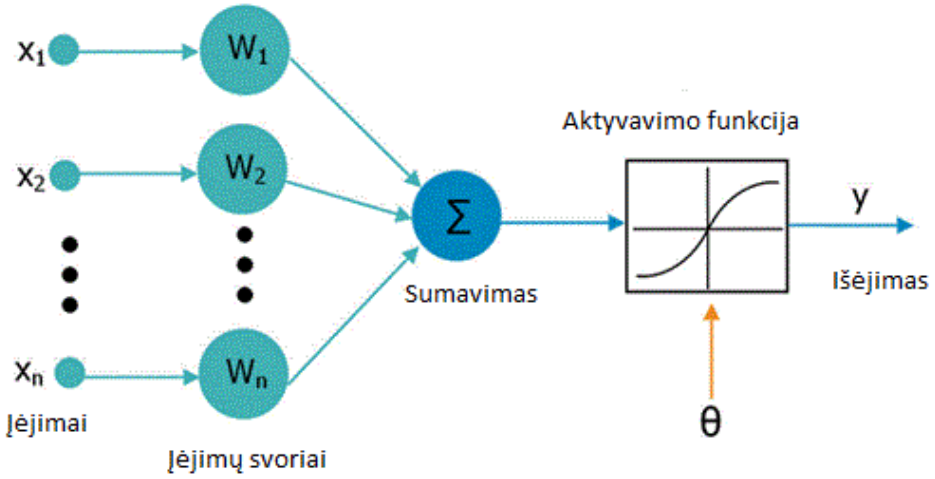
Pats dirbtinis neuronas yra sudarytas iš trijų pagrindinių komponentų: svoriai, slenksčiai ir aktyvavimo funkcija (žr. 1pav.). Svorių koeficientai $w_1, w_2, ..., w_n$ rodo stiprumus atskirų įvesčių, aprašytų vektoriumi $x_1, x_2, ..., x_n$. Neuroninė jungtis apskaičiuojama, kiekvieną įvesties signalą dauginant iš svorio koeficiento. Gauname, kad kiekvienas neurono tinklo išėjimas yra apskaičiuojamas pagal tokią formulę [22]:

$$y = f\left(\sum_{i=1}^{n} x_i w_i\right); \tag{5}$$

čia y – išėjimas, f – aktyvavimo funkcija, n – prieš tai buvusio sluoksnio neuronų skaičius, x_i – įėjimai ir w_i – įėjimų svoriai.

Jei svorio koeficiento reikšmė yra teigiama, tada sužadina signalą išvestyje, kai svorio koeficiento reikšmė neigiama, slopina išvesties signalą.

Aktyvavimo funkcija yra matematinė operacija kartu su išvesties signalu, kurią naudojant transformuojamas gautas rezultatas (žr.1 pav.). Nuo dirbtinio neuroninio tinklo sprendžiamo uždavinio priklauso, kokio sudėtingumo aktyvavimo funkcija taikoma. Tarp dažniausiai naudojamų yra tiesinė, slenksčio, sigmoidinė, hiperbolinė tangento, logistinė ir Piesewise‘o aktyvavimo funkcijos [19].



1 pav. Dirbtinio neuroninio tinklo sandara

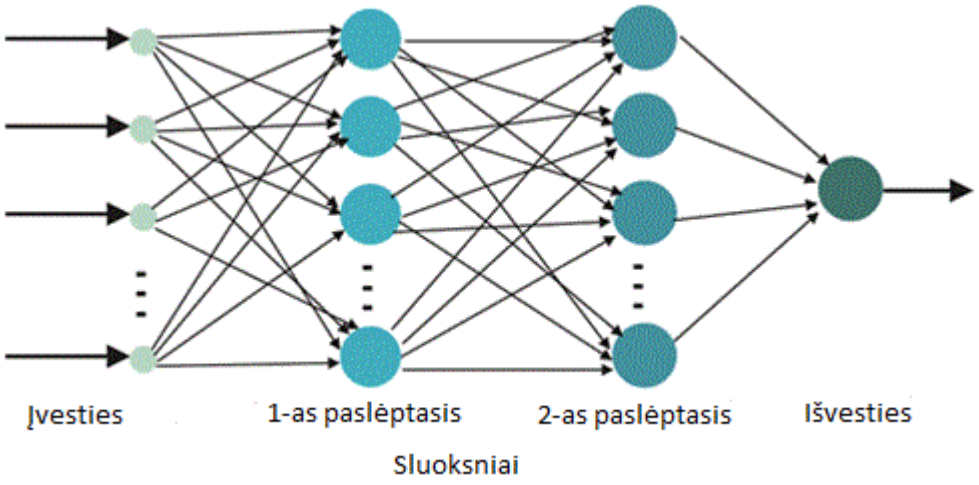
Išvesties aktyvavimą vidinis neurono slenkstis veikia taip [22]:

$$y = f\left(\sum_{i=1}^{n} x_i w_i - \theta\right), \tag{6}$$

čia θ – dirbtinio neurono slenkstis.

Yra skiriami dviejų tipų neuroniniai tinklai, vienasluoksnis ir daugiasluoksnis perceptronai:

- Vienasluoksnis perceptronas – tarp įėjimo ir išėjimo sluoksnių nėra paslėptų neuronų sluoksnių (žr. 1 pav.);
- Daugiasluoksnis perceptronas – tarp įėjimo ir išėjimo sluoksnių yra viena ir daugiau paslėptų neuronų sluoksnių (žr. 2 pav.);



2 pav. Daugiasluoksnis perceptronas

Vieno sluoksnio dirbtinių neuroninių tinklų (vieno perceptrono) nepakanka spręsti problemai, kuri turi netiesinį sprendimą. Todėl tokiam uždaviniui spręsti, naudojamas neuroninių tinklų modelis su

vienu ar daugiau paslėptųjų sluoksnių, kai išvedimas perduodamas iš vieno perceptrono į kitą. Informacija daugiasluoksniame perceptrone perduodama viena kryptimi, iš įvesties sluoksnio į išvesties sluoksnį (žr. 2 pav.).

Kad neuroninis tinklas galėtų išspręsti uždavinius, jis turi būti apmokomas iš turimų pavyzdžių. Metodo išskirtinumas tas, kad jis apsimoko pats ir jam nereikia išankstinių specifikacijų ar formulių, o vietoje parametrų nustatymų, naudojami išoriniai duomenys. Apmokymas vykdomas, keičiant tarp neuronų esančių jungčių svorius. Taigi neuroniniai tinklai ieško geriausio uždaviniui tinkančio sprendimo.

Dirbtiniai neuroniniai tinklai pasižymi šiomis charakteristikomis [23, 24]:

- Saviorganizacija ir adaptyvumas – metodui būdinga saviorganizacija, be to, modelis lengvai adaptuojasi prie duomenų pasikeitimų, o visa tai lemia galingas ir efektyvias duomenų apdorojimo galimybes.
- Lygiagretūs skaičiavimai – skaičiavimai yra atliekami tarp daugybės tarpusavyje susietų neuronų.
- Netiesinis tinklo apdorojimas – padidina neuroninio tinklo apibendrinimo laipsnį, galimybes pašalinti triukšmą duomenyse, bei geresnį klasifikavimą.
- Tolerancija klaidoms – jei dalis neuroninio tinklo netinkamai veikia, sumažėja metodo efektyvumas ir tikslumas, vis dėl to, tam tikros neuroninio tinklo funkcijos vis vien gali būti patikimai vykdomos.

Sprendžiant klasifikavimo uždavinį neuroninių tinklų metodu, įvesties sluoksnyje pateikiami požymius aprašantys duomenys, o išvesties sluoksnyje yra gaunamas rezultatas nusakantis prognozuojamo kintamojo priklausymą klasėms.

Šiame darbe atsiliepimų naudingumui klasifikuoti naudojamas R programinio paketo funkcija *nnet* iš *nnet* paketo. Ši funkcija pagal nutylėjimą taiko dirbtinius neuroninius tinklus su vienu paslėptuoju sluoksniu [25].

2.4.2. Atsitiktinių miškų metodas

Atsitiktinių miškų metodas (angl. *Random Forests*) naudojamas klasifikavimo uždaviniams spręsti. Vykdant metodą pradinė mėginių aibė *M* atsitiktiniai padalinama į *S* poaibių, du trečdaliai poaibių naudojami apmokymui, o likę OOB (*Out of Bag*) testavimui. Iš poaibių suformuojama *S* pilnų sprendimų medžių, kurie ir sudaro atsitiktinį mišką. Testuojama mėginių aibė klasifikuojama su visais sprendimų medžiais. Kiekvienas sprendimų medis mėginį priskiria tam tikrai klasei. Po klasifikacijos, remiantis gautų klasių pasikartojimų dažniu, mėginys priskiriamas tai klasei, kurios

pasikartojimų dažnis buvo didžiausias. Kad atrinkti tinkamiausią mėginių aibės dalinimui naudojamą požymį, skaičiuojamas informacijos kiekis, tarp kiekvieno požymio *A* ir klasių iš aibės *M*. Požymių svarbumui apskaičiuoti naudojami OOB duomenys. Informacijos kiekio įvertis parodo, kaip gerai parinktas požymis padalina mėginių aibę į skirtingas klases. Parenkamas požymis tas, kuris turi didžiausią informacijos kiekio įvertį. Informacijos kiekis apskaičiuojamas pagal formulę:

$$Gain(M, A) = H(M) - sum\left(\left(\frac{|S|}{M}\right) \times H(S)\right), \tag{7}$$

čia *M* – mėginių aibė, *A* – požymis, S – mėginių aibės poaibis, H(M) – entropija, kuri apskaičiuojama:

$$H(M) = -p(+) \times \log 2(p(+)) - p(-) \times \log 2(p(-)); \tag{8}$$

kur *p*(-) yra tikimybė, kad mėginys priklauso tam tikrai klasei.

Spendimų medžiai yra lengvai interpretuojami, gali būtini naudojami su nepilnais duomenimis (angl. *Missing data*), gerai susitvarko su triukšmu, bei gana greitai veikia, taip pat klasifikavimą atlieką gerai ir su maža duomenų imtimi. Tačiau sprendimų medžių metodas linkęs persimokyti (angl. *overfitting*), tai yra su nauja duomenų imti modelis skaičiuoja blogai, nors su mokymosi duomenimis buvo gautas geras rezultatas. Tai pat, kuo gilesnis ir kuo didesnis medis tuo daugiau kompiuterio atminties naudojama.

2.4.3. Artimiausių kaimynų metodas

Artimiausių kaimynų (angl. *k-nearest neighbors*) metodas priskiriamas klasifikavimo be mokytojo metodams [26]. Jo idėja palyginti naują objektą su panašiais į jį mokymosi aibės objektais. Kad objektą priskirti kuriai nors klasei, skaičiuojami atstumai nuo objekto iki visų mokymosi aibės objektų. Atstumui išmatuoti dažniausiai naudojamas Euklido $d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$, Čebyševo $Max = (|x - y|)$ arba miesto kvartalo $|x - y|$ atstumas [27].

Tikslumas metodo priklauso nuo parinkto kaimynų skaičiaus *k* kuris parenkamas eksperimentiškai. Naujas objektas priskiriamas klasei, kurioje yra dauguma jo kaimynų iš artimiausių *k*. Taigi artimiausių kaimynų metodo žingsniai yra tokie [27]:

- parenkamas artimiausių kaimynų skaičius *k;*
- apskaičiuojami atstumai tarp naujo elemento ir jau išskirstytų ar įvestų apmokymo elementų;
- surūšiavus objektus pagal atstumą, atrenkami *k* arčiausiai esančių objektų;

- arčiausiai esantys *k* kaimynai yra priskiriami sričiai;
- naujam objektui parenkama sritis, turinti daugiausiai atrinktų kaimynų .

Šio modelio trūkumas, tas, kad kuo didesnis *k* kaimynų skaičius parenkamas, tuo modelis didesnis, o tai ilgina skaičiavimo laiką.

2.5. Klasifikavimo modelių tinkamumo duomenims vertinimas

2.5.1. Kryžminis patikrinimas

Klasifikavimo metodų apmokymui ir testavimui naudojamas blokų kryžminio patikrinimo metodas (angl. *k – fold cross validation*). Metodu duomenų aibė suskaidoma į *k* nesikertančių imčių. Algoritmų mokymasis vyksta naudojant *k*–1 imties duomenis, o su likusia duomenų dalimi algoritmas testuojamas fiksuojant klasifikavimo matų reikšmes. Procedūra kartojama *q* kartų ir imant vis kitas *k*–1 imtis. Pabaigoje imamos klasifikavimo rezultatų vidutinės reikšmės. Tyrime pasirinktas *k* imčių skaičius yra 3, tai yra duomenys suskaidoma į tris nesikertančias dalis.

2.5.2. Sumaišymo matrica

Pasitikrinimui ar modelis gerai klasifikuoja, naudojama sumaišymo matrica (angl. *Confusion matrix*), kuri parodo teisingai ir neteisingai suklasifikuotų duomenų atvejų skaičių. Remiantis matricos rezultatais galima nustatyti klasifikatoriaus klaidas, bei įvertinti jo tikslumą. Sumaišymo matricos schema pateikta 3 lentelėje.

3 lentelė. Sumaišymo matrica

		Tikrosios klasės	
		Naudingų atsiliepimų klasė	Nenaudingų atsiliepimų klasė
Prognozuojamos klasės	Naudingų atsiliepimų klasė	TP	FP
	Nenaudingų atsiliepimų klasė	FN	TN
Viso atvejų		P	N

TP (angl. *True positive*) – naudingi atsiliepimai, kurie priskirti naudingų atsiliepimų klasei, kai iš tiesų jie šiai klasei ir priklauso;

FP – nenaudingi atsiliepimai, kurie priskirti naudingų atsiliepimų klasei, kai iš tiesų jie šiai klasei nepriklauso;

FN – naudingi atsiliepimai, kurie priskirti nenaudingų atsiliepimų klasei, kai iš tiesų jie šiai klasei nepriklauso;

TN – nenaudingi atsiliepimai, kurie priskirti nenaudingų atsiliepimų klasei, kai iš tiesų jie šiai klasei ir priklauso;

N –bendras nenaudingų atsiliepimų klasės skaičius;

P – bendras naudingų atsiliepimų klasės skaičius.

Klasifikatoriaus klaidos yra FP ir FN kurios parodo, kad tiek atvejų buvo suklasifikuota neteisingai, o TP ir TN yra gerai suklasifikuoti atvejai.

Klasifikavimo kokybei įvertinti dažniausiai skaičiuojamas specifiškumas, jautrumas ir bendras klasifikavimo tikslumas [26]:

$$jautrumas = \frac{TP\ skaičius}{TP\ skaičius + FN\ skaičius}; \tag{7}$$

$$specifiškumas = \frac{TN\ skaičius}{TN\ skaičius + FP\ skaičius}; \tag{8}$$

$$bendras\ klasifikavimo\ tikslumas = \frac{TP\ skaičius + TN\ skaičius}{N + P}. \tag{9}$$

Jautrumas parodo teigiamų, gerai suklasifikuotų įvykių dalį, o specifiškumas neigiamų gerai suklasifikuotų įvykių dalį. Bendras klasifikavimo tikslumas, tai dalis gerai modeliu suklasifikuotų atvejų. Visos trys charakteristikos gali būti išreikštos procentais.

2.5.3. ROC ir DET kreivės

Klasifikavimo metodų, rizikos vertinimo jautrumui ir specifiškumui palyginti naudojamos kreivės ROC (angl. *Receiver Operating Characteristics*) ir DET (angl. *detection error trade-of*). Braižant ROC kreivę, X ašyje atidedamas teisingai klasifikuotų nenaudingų atsiliepimų santykis (specifiškumas), o Y ašyje teisingai klasifikuotų naudingų atsiliepimų santykis (jautrumas). Apskaičiuotas plotas po ROC kreive (Area Under the Curve, AUC), įvertinamas AUC verte. Jeigu AUC yra mažiau kaip 0,5, klasifikatorius nepasižymi prognostinėmis savybėmis, jei AUC vertė daugiau kaip 0,7, klasifikatorius gali būti naudojamas prognozavimui [28].

Kreivės DET Y ašyje atidedamas klaidingai klasifikuotų nenaudingų atsiliepimų procentas (neteisingai neigiama klaida), o X ašyje klaidingai priskirtų naudingų atsiliepimų procentas (neteisingai teigiama klaida). Kreivės DET taškas, kuriame specifiškumas lygus jautrumui yra vadinamas lygio klaidų tašku, kuris žymimas EER. Kuo EER mažesnis tuo klasifikavimas tikslesnis.

3. Tyrimo rezultatai ir jų aptarimas

Šiame skyriuje pateikiami realių vartotojų atsiliepimų internete tyrimo rezultatai.

3.1. Tyrimo duomenys

Darbe naudojami *Yelp.com* svetainės 2017 metų duomenų rinkinys, patalpintas puslapyje https://www.kaggle.com/yelp-dataset/yelp-dataset/data. Visi failai pateikti csv formatu. Duomenų rinkinį sudaro 5,2 milijonų atsiliepimų, 174 tūkstančiai verslo objektų. Panaudoti du failai yelp_business.csv ir yelp_review.csv. Faile yelp_business.csv yra informacija susijusi su paslaugomis ir vartotojų įvertinimais, kitame duomenys apie vartotojų atsiliepimus. Duomenų filtravimo etapai pateikti 4 lentelėje. Iš yelp_review.csv duomenų failo, kuriame yra informacija apie atsiliepimus, nuskaitoma 400000 atsiliepimų. Tada yelp_business.csv faile duomenys filtruojami pagal verslo kategoriją, apie restoranų kategorijai priskiriamus verslo objektus (toliau vadinami „restoranais“). Po filtracijos duomenų failą sudaro 54630 skirtingų restoranų duomenys. Sekančiame žingsnyje, minėti failai sujungiami pagal sutampančias *business_id* stulpelio reikšmes. Po šio žingsnio duomenų failą sudaro 241284 eilutės. Be to, filtruojami trijų metų atsiliepimai nuo 2015 m. sausio 1 dienos iki naujausio duomenų faile esančio atsiliepimo (2017 metų gruodžio 11 dienos), taip pat atsiliepimai kurių naudingumas yra ne mažesnis už 4 atsiliepimo naudingumo balsus.

4 lentelė. Duomenų failo filtravimo etapai

Failo filtravimo etapai	**Atsiliepimų skaičius duomenų faile pieš filtravimą**	**Filtravimo priežastis**	**Atsiliepimų skaičius duomenų faile po filtravimo**
1 etapas	400000	Restoranų kategorijai priklausantys atsiliepimai	241284
2 etapas	241284	Atsiliepimai kurių data nuo 2015 metų sausio 1 dienos	140947
3 etapas	140947	Atsiliepimai kurių naudingumas ne mažesnis už 4 naudingumo balsus	7023

Sutvarkytame duomenų faile pateikti 7023 atsiliepimai, apie 4677 skirtingus restoranus, juos parašė 3036 skirtingi autoriai (vartotojai) (žr. 5 lentelę).

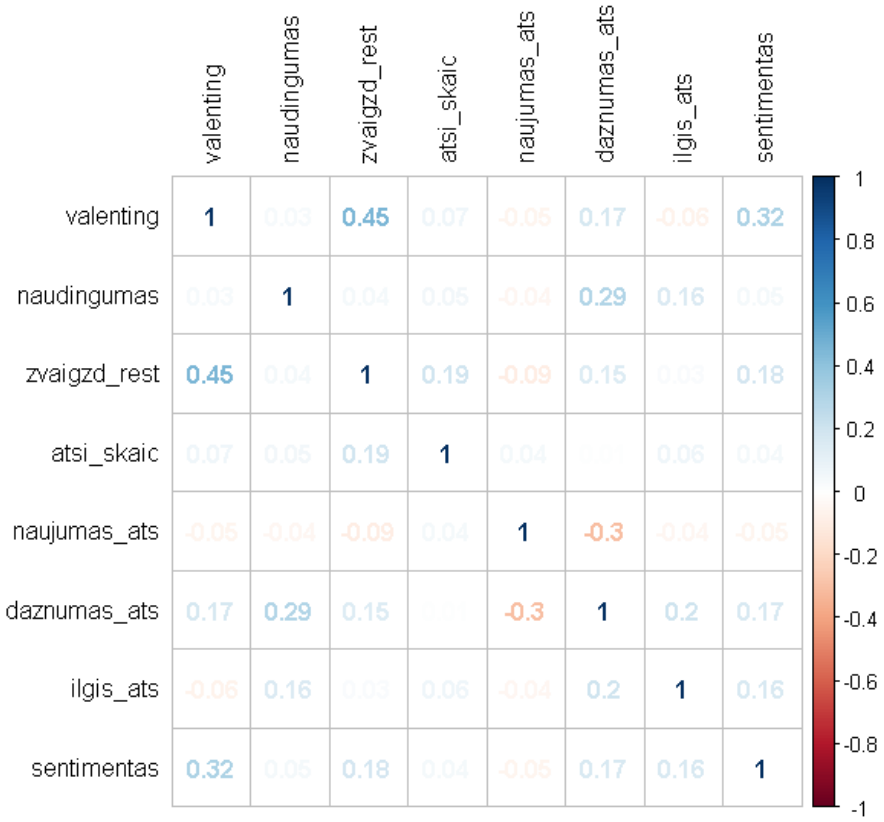
5 lentelė. Duomenų imties charakteristikos

Naujausias atsiliepimas	Seniausias atsiliepimas	Restoranų skaičius	Atsiliepimų autorių skaičius
11-12-2017	01-01-2015	4677	3036

Sudaryta duomenų matrica, kuri naudojama taikant regresinę analizę, bei prognozuojant atsiliepimų naudingumą. Šią matricą sudaro priklausomas kintamasis Y (naudingumas), bei kiti kintamieji (veiksniai), apie juos daugiau informacijos pateikta 2.1 ir 2.2 skyreliuose.

3.2. Neigiamos binominės regresinės analizės rezultatai

Koreliacija tarp kintamųjų įvertinta taikant Spirmeno koreliacijos koeficientą, nes kintamųjų skirstiniai nėra normalieji. 3 pav. matome, kad multikolinearumo tarp nepriklausomų kintamųjų nėra, o su priklausomu kintamuoju *naudingumas*, kiti kintamieji neturi stipraus ryšio.



3 pav. Kintamųjų koreliacijos

Siekiant įvertinti, kurie analizuojami veiksniai yra statistiškai reikšmingi atsiliepimų naudingumui, taikoma neigiama binominė regresija. Sudarytas modelis yra geras, kai tenkinamos kelios prielaidos. Viena jų, priklausomo kintamojo dispersija turi būti didesnė už vidurkį. Kadangi priklausomo kintamojo atsiliepimo naudingumas dispersija lygi 22,42 ir ji yra didesnė už vidurkį 7,59, prielaida yra tenkinama (žr. 6 lentelę). Sudaryto regresijos modelio AIC reikšmė yra lyginama su regresijos modelio AIC reikšme, kuriame yra tik konstanta (žr. 6 lentelė).

6 lentelė. Kintamojo naudingumas statistika

Vidurkis	Dispersija	AIC (modelio tik su konstanta)
7.5939	22.4196	41230

Neigiamos binominės regresijos (žr. 7 lentelę) kintamieji *valentingumas*, *neigiamas sentimentas*, *teigiamas sentimentas*, nėra statistiškai reikšmingi ($p > 0,05$) atsiliepimų *naudingumo* vidurkiui, todėl modelis buvo tobulinamas.

7 lentelė. Neigiamos binominės regresijos modelis su visais kintamaisiais

Regresijos lygties koeficientas	**Regresijos lygties kintamasis**	**Regresijos lygties koeficiento taškinis įvertis**	**p-reikšmė p>|z|**
$\hat{\beta}_0$	*konstanta*	1.559e+00	< 2e-16 ***
$\hat{\beta}_1$	*valentingumas*	3.694e-03	0.584
$\hat{\beta}_3$	*restorano vertinimas žvaigždute*	1.083e-01	2.13e-11 ***
$\hat{\beta}_4$	*naujumas*	-2.946e-04	< 2e-16***
$\hat{\beta}_5$	*atsiliepimų skaičius.*	6.687e-05	4.69e-10 ***
$\hat{\beta}_6$	*ilgis*	4.886e-04	< 2e-16***
$\hat{\beta}_7$	*dažnumas*	5.887e-03	< 2e-16***
$\hat{\beta}_8$	*neigiamas sentimentas*	5.126e-02	0.382
$\hat{\beta}_9$	*teigiamas sentimentas*	6.393e-02	0.198

Reikšmingumo lygmuo 0,001 ***, 0,01**, 0.05 *, 0,1 .

Kituose regresijos modeliuose, į lygtį buvo įtrauktos visų kintamųjų sąveikos su sentimentų pseudokintamaisias (pvz. *naujumas * teigiamas sentimentas, naujumas* neigiamas sentimentas*, *ilgis * teigiamas sentimentas*, *ilgis* neigiamas sentimentas* ir taip toliau). Sudarytos dvi regresijos lygtys, kurios išpildo gero modelio reikalavimus, vienas iš jų, kad visi lygties kintamieji (regresoriai) turi būti statistiškai reikšmingi (žr. 8 ir 10 lenteles).

Deviacijos ir jos laisvės laipsnio santykis 1 modelyje lygus 0,99 ir jis artimas vienetui, tai parodo, kad modelis nedaug skiriasi nuo visiškai duomenis aprašančio modelio, taigi modelis gerai tinka duomenims. AIC reikšmė 40752 yra mažesnė už modelio tik su konstanta 41230, todėl patvirtinama dar viena modelio tinkamumo prielaida. Kad sužinoti ar modelyje yra bent vienas *naudingumui* prognozuoti reikalingas nepriklausomas kintamasis, tikrinama didžiausio tikėtinumo santykio chi kvadrato statistika, kurios stebėta reikšmė 488,37, o $p = 0,00$ ir yra mažesnė už 0,05 reikšmingumo lygmenį, taigi patvirtinama, kad modelyje yra bent vienas reikšmingas regresorius kintamojo prognozavimui. Remiantis Voldo kriterijaus *p* reikšmėmis, kurios pateiktos 9 lentelėje, matome,

kad statistiškai reikšmingi visi modelio kintamieji ($p < 0,05$). Taigi 1 modelį sudaro *restorano vertinimas žvaigždute*, *atsiliepimų skaičius*, *ilgis* ir *dažnumas*.

8 lentelė. Neigiamos binominės regresijos 1 modelis

Regresijos lygties koeficientas	**Regresijos lygties kintamasis**	**Regresijos lygties koeficiento taškinis įvertis**	**p-reikšmė p>|z|**	**Akaikės informacinis kriterijus (AIC)**	**Deviacijos ir laisvės laipsnių santykis**
$\hat{\beta}_0$	*konstanta*	1.321e+00	< 2e-16 ***	40752	0.9939
$\hat{\beta}_1$	*restorano vertinimas žvaigždute*	1.266e-01	< 2e-16 ***		
$\hat{\beta}_2$	*atsiliepimų skaičius*	6.773e-05	3.8e-10 ***		
$\hat{\beta}_3$	*ilgis*	4.914e-04	< 2e-16 ***		
$\hat{\beta}_4$	*dažnumas*	7.739e-03	< 2e-16 ***		

Reikšmingumo lygmuo 0,001 ***, 0,01**, 0.05 *, 0,1 .

1 modelio regresijos lygtis:

$$Naudingumo\_vidurkis = \exp(\hat{\beta}_0 + \hat{\beta}_1 \cdot restorano\_žvaigžd\_įv. + \hat{\beta}_2 \cdot atsiliepimų\_skaič + \\ + \hat{\beta}_3 \cdot ilgis + \hat{\beta}_4 \cdot dažnumas). \quad (9)$$

Modelio koeficientų įverčių eksponentės pateiktos 9 lentelėje. Visos regresijos koeficientų reikšmės yra teigiamos (žr. 8 lentelę), todėl didėjant kintamųjų *restorano vertinimas žvaigždute, atsiliepimų skaičius*, *ilgis* ir *dažnumas* reikšmėms, didėja ir atsiliepimų *naudingumo* vidurkis. Restoranų įvertintų viena papildoma žvaigždute atsiliepimo naudingumo vidurkis didesnis 1,14 karto. Kiekvienas papildomas atsiliepimas apie restoraną, didina 1,00007 karto panašias kintamųjų reikšmes, turinčių restoranų atsiliepimų *naudingumo* vidurkį. Atsiliepimo tekste, kiekvienas papildomas žodis atsiliepimų *naudingumo* vidurkį padidina 1,0005 karto, o vienas vartotojo papildomai parašytas atsiliepimas, padidina atsiliepimų naudingumo vidurkį 1,0078 karto.

9 lentelė. Neigiamos binominės regresijos 1 modelio koeficientų įverčių eksponentės

Regresijos lygties koeficientas	**Regresijos lygties kintamasis**	**Koeficientų įverčių eksponentės**
$\hat{\beta}_0$	*konstanta*	3.748267
$\hat{\beta}_1$	*restorano vertinimas žvaigždute*	1.134965
$\hat{\beta}_3$	*atsiliepimų skaičius*	1.000068
$\hat{\beta}_4$	*ilgis*	1.000492
$\hat{\beta}_5$	*dažnumas*	1.007769

Deviacijos ir jos laisvės laipsnio santykis 2 modelyje lygus 0,92, kadangi jis yra intervale nuo 0,9 iki 1,1, todėl modelis nedaug skiriasi nuo visiškai duomenis aprašančio modelio (žr. 10 lentelę). AIC reikšmė 40824 yra mažesnė už modelio tik su konstanta reikšmę 41230, todėl patvirtinama dar viena modelio tinkamumo prielaida duomenims. Didžiausio tikėtinumo santykio chi kvadrato statistika lygi 416,16, o p = 0,00, tai patvirtina, kad modelyje yra bent vienas naudingumo prognozavimui reikšmingas kintamasis. Remiantis Voldo kriterijau p reikšmėmis, kurios pateiktos 10 lentelėje, nustatyti statistiškai reikšmingi veiksniai: *atsiliepimų skaičius*, *ilgis* ir *dažnumas* kintamieji ($p < 0,05$). *Neigiamo sentimento* $p < 0,05$, taigi su 95 procentų garantija laikoma, kad *neigiamas sentimentas* statistiškai reikšmingas atsiliepimų *naudingumo* vidurkiui. 2 modelį sudaro, *atsiliepimų skaičius*, *ilgis*, *dažnumas*, bei *neigiamas sentimentas*.

10 lentelė. Neigiamos binominės regresijos 2 modelis.

Regresijos lygties koeficientas	**Regresijos lygties kintamasis**	**Regresijos lygties koeficiento taškinis įvertis**	**p-reikšmė p>|z|**	**Akaikės informacinis kriterijus (AIC)**	**Deviacijos ir laisvės laipsnių santykis**
$\hat{\beta}_0$	*konstanta*	1.801e+00	< 2e-16 ***	40824	0.9248
$\hat{\beta}_1$	*atsiliepimų skaičius*	7.826e-05	3.87e-13 ***		
$\hat{\beta}_3$	*ilgis*	4.972e-04	3.8e-10 ***		
$\hat{\beta}_4$	*dažnumas*	8.186e-03	< 2e-16 ***		
$\hat{\beta}_5$	*neigiamas sentimentas*	-7.384e-02	0.036 *		

Reikšmingumo lygmuo 0,001 ***, 0,01**, 0.05 *, 0,1 .

2 modelio regresijos lygtis :

$$Naudingumo\_vidurkis = \exp(\hat{\beta}_0 + \hat{\beta}_1 \cdot atsiliepimų\_skaič + \hat{\beta}_2 \cdot ilgis + \\ + \hat{\beta}_3 \cdot dažnumas + \hat{\beta}_4 \cdot neigiamas\_sentimentas) \quad (11)$$

Koeficientų $\hat{\beta}_i$ reikšmės yra 10 lentelėje.

Antro modelio koeficientų įverčių eksponenčių reikšmės pateiktos 11 lentelėje. Prie visų kintamųjų išskyrus *neigiamas sentimentas*, regresijos koeficientai yra teigiami (žr. 10 lentelę), todėl didėjančios kintamųjų *atsiliepimų skaičius*, *ilgis* ir *dažnumas* reikšmės didina atsiliepimų *naudingumo* vidurkį. Kitas regresijos lygties rezultatas parodo, kad *neigiamą sentimentą* turinčių atsiliepimų naudingumo vidurkis yra vidutiniškai 0,93 karto mažesnis, lyginant su neutralių ir teigiamų atsiliepimų naudingumo vidurkiu. Kiekvienas papildomai parašytas autoriaus atsiliepimas, naudingumo vidurkį padidina 1,0082 karto.

11 lentelė. Neigiamos binominės regresijos 2 modelio koeficientų įverčių eksponentės

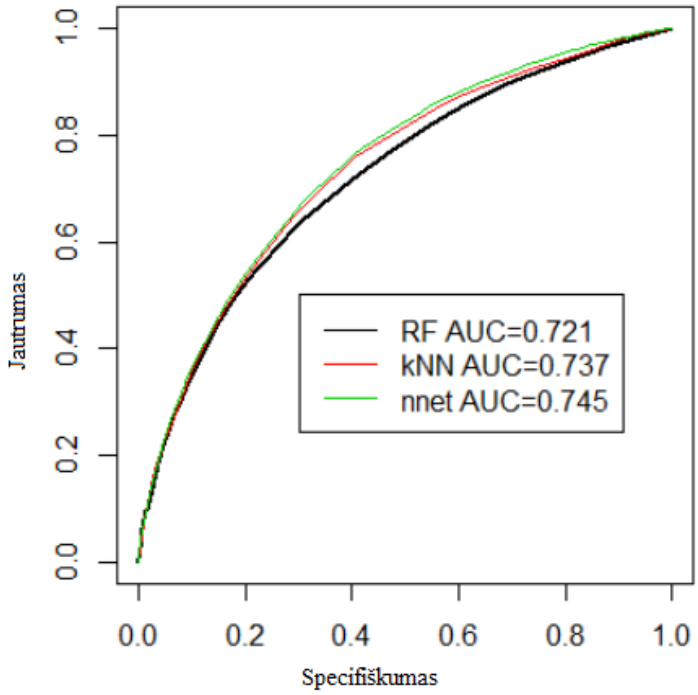
Regresijos lygties koeficientas	**Regresijos lygties kintamasis**	**Koeficientų įverčių eksponentės**
$\hat{\beta}_0$	*konstanta*	6.0543952
$\hat{\beta}_1$	*atsiliepimų skaičius*	1.0000783
$\hat{\beta}_3$	*ilgis*	1.0004973
$\hat{\beta}_4$	*dažnumas*	1.0082196
$\hat{\beta}_5$	*neigiamas sentimentas*	0.9288246

Remiantis Akaikės informaciniu kriterijumi (AIC) ir deviacijos bei jos laisvės laipsnių santykiu, nežymiai, bet geresnis yra pirmasis modelis, su nepriklausomais kintamaisiais *restorano vertinimas žvaigždute, atsiliepimų skaičius*, *ilgis* ir *dažnumas*.

3.3. Atsiliepimų internete klasifikavimo modelių taikymo rezultatai

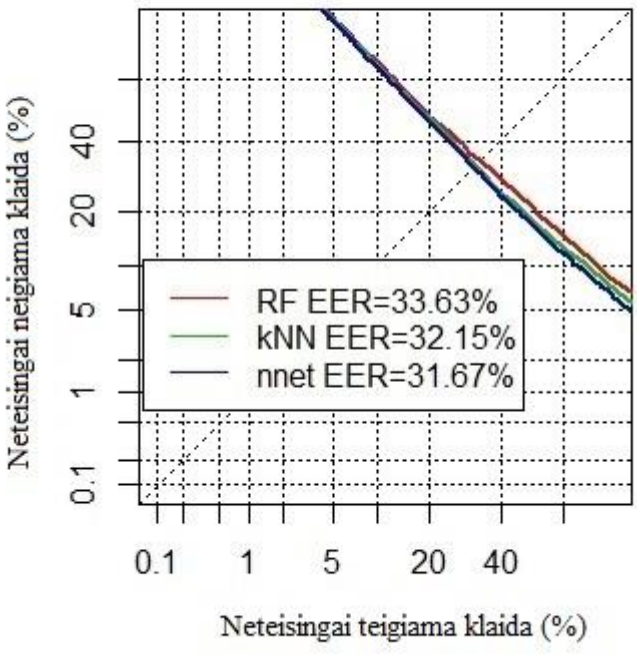
Atsiliepimų klasifikavimui į naudingus ir nenaudingus, buvo naudojama 35559 atsiliepimų imtis. Naudingais atsiliepimais buvo tie, kurie turi ne mažiau kaip du naudingumo balsus. Duomenų imtį sudarė 6042 naudingi atsiliepimai ir 29517 nenaudingų. Klasifikavimo uždavinys sprendžiamas neuroninių tinklų (*Neural Network*), artimiausių kaimynų (angl. *k-Nearest Neighbors*) ir atsitiktinių miškų (angl. *Random Forest*) metodais. Be to, pritaikytas kryžminis patikrinimas (angl. *k – fold cross validation*), kai $k = 3$.

Atsiliepimų klasifikavimas realizuojamas remiantis neigiamos binominės regresinės analizės rezultatais. Naudojami pirmojo regresijos modelio kintamieji *atsiliepimų skaičius*, *ilgis*, *dažnumas*, *restorano vertinimas žvaigždute*. Iš ROC kreivės (žr. 4 pav.) galima daryti išvadą, kad nežymiai, bet geriau už artimiausių kaimynų (AUC reikšmė lygi 0,737) ir atsitiktinių miškų (AUC reikšmė lygi 0,721) metodus klasifikuoja neuroniniai tinklai, kadangi tada AUC (plotas po kreive) yra didžiausias 0,745. Sudarytas neuroninis tinklas yra su vienu paslėptuoju sluoksniu, kuriame yra 10 neuronų.



4 pav. ROC kreivė (plotas po kreive AUC)

Nors DET kreivės lygių klaidų taško EER reikšmės kiekvieno modelio ne daug skiriasi, vis dėl to mažiausia 31,67 % reikšmė klasifikuojant atsiliepimus neuroninių tinklų metodu (žr. 5 pav.).



5 pav. DET kreivė (lygių klaidų lygis EER)

Bendras prognozavimo tikslumas lygus 68,48 %. Jautrumas lygus 67.81 %, tiek procentų naudingų atsiliepimų buvo teisingai priskirta naudingų atsiliepimų klasei. Klasifikavimo specifiškumas lygus 68,62 %, tiek procentų nenaudingų atsiliepimų neuroninių tinklų metodu, buvo gerai suklasifikuoti. Kadangi tyrime svarbu yra teisingai atskirti naudingus atsiliepimus, todėl reikia atkreipti dėmesį, į tai kad, naudingi atsiliepimai sudarė 17% duomenų imties, o į tai atsižvelgus, jautrumas rodo pakankamai gerą naudingų atsiliepimų klasifikavimo rezultatą.

12 lentelė. Neuroninių tinklų algoritmo rezultatų sumaišymo matrica

		Tikros reikšmės		**Visų prognozuojamų atvejų skaičius**
		Naudingų atsiliepimų klasė	**Nenaudingų atsiliepimų klasė**	
Prognozės rezultatas	**Naudingų atsiliepimų klasė**	4097	9262	13359
	Nenaudingų atsiliepimų klasė	1945	20255	22200
Visų tikrų atvejų skaičius		6042	29517	35559
Prognozavimo tikslumas proc.		**Jautrumas (proc.)**	**Specifiškumas (proc.)**	**Bendras klasifikavimo tikslumas (proc.)**
		67.81	68.62	68.48

Nors tarp dviejų klasifikuojamų klasių duomenų imtyje yra disbalansas, (naudingi atsiliepimai sudaro 17% duomenų imties, o nenaudingi 83%) dirbtinių neuroninių tinklų metodu pavyko tiek naudingų, tiek nenaudingų atsiliepimų klases suklasifikuosi panašiu tikslumu.

3.4. Tyrimo rezultatų apibendrinimas ir diskusija

Vartotojų internete ilgesni atsiliepimai, kaip ir kituose tyrimuose [1,2,7], nustatyta kad teigiamai veikia atsiliepimų naudingumą, todėl galima manyti, kad ilgesniuose tekstuose pateikiama daugiau naudingos informacijos. Tyrimo rezultatai patvirtino ir kituose moksliniuose darbuose [8] gautas išvadas, kad bendras visų vartotojų įvertinimas žvaigždute yra vienas stipriausiai veikiančių atsiliepimų naudingumą rodiklių. Atsiliepimai, kuriuose kalbama apie gerai įvertintą restoraną yra naudingesni. Vartotojo patirtis ir įžvalgos apie restoraną, pateikiamos internetinėje svetainėje atsiliepimo forma, bei restorano įvertinimu žvaigždute (skalėje nuo 1 iki 5), o jis priklauso nuo to, kokią apsilankymo restorane patirtį vartotojas turėjo. Todėl galima manyti, kad geri atsiliepimai apie restoraną, kurie tuo pačiu gali būti ir naudingi, sąlygoja gerą restorano įvertinimą žvaigždute.

Tai pat šio tyrimo rezultatai parodė, kad didesnį poveikį lyginant su kitais veiksniais, atsiliepimo internete naudingumui daro neigiamas atsiliepimo sentimentas. Nors neigiami atsiliepimai gali būti tokie pat naudingi kaip ir atsiliepimai, kuriuose parašyta teigiama informacija apie restoraną, tačiau tokius rezultatus galėjo lemti tai, kad neigiamuose atsiliepimuose apie restoranus nėra pateikiama

svarbi informacija, kuri galėtų būti reikšminga didesnei atsiliepimus skaitančiai vartotojų auditorijai.

Kiti šiame tyrime nustatyti vartotojų atsiliepimų internete naudingumui įtaką darantys veiksniai yra atsiliepimų skaičius ir dažnumas. Tyrimo rezultatai rodo, kad augantis internetinėje svetainėje parašytų atsiliepimų skaičius apie restoraną, siejamas su didesniu atsiliepimų naudingumu. Didelis atsiliepimų skaičius gali parodyti, kad restoranas yra populiarus, o dažniausiai populiarumas siejamas su teigiama nuomone, kuria dalinamasi internetinėse svetainėse. Kuo daugiau vartotojų skaito atsiliepimus, tuo didesnė tikimybė, kad bus įvertinta daugiau atsiliepimų, o rezultate ir daugiau atsiliepimų, kurie surinktų didesnį naudingumo balsų skaičių. Atsiliepimų dažnumas nustatyta tai pat teigiamai veikia naudingumą. Vartotojas kuris ankščiau parašė nemažai atsiliepimų gali būti, kad yra patikimesnis ir turintis daugiau patirties, todėl ir jo atsiliepimai yra laikomi naudingesniais.

Šiame poskyryje pateiktas įžvalgas reikia patikrinti detaliau. Galima manyti, kad atsiliepimų naudingumas priklauso nuo restorano populiarumo, bei gebėjimo sudominti vartotoją. Be to, vartotojai kurie aktyviai įsitraukia internetinėje svetainėje vertinant ir reiškiant nuomonę apie restoranus, gali geriau atskirti kokia informacija yra naudinga, o kokia ne. Taigi, tokie vartotojai nereikšmingos informacijos savo atsiliepimuose stengiasi nepateikti.

3.5. Atsiliepimų internete klasifikavimo modelio taikymas

Darbe sukurtas atsiliepimų klasifikavimo modelis, internetinėse svetainėse gali automatiškai klasifikuoti atsiliepimus. Naudingų atsiliepimų publikavimas dažniausiai lankomose internetinės svetainės vietose, pritrauktų daugiau lankytojų. Modelis gali būti pritaikomas bet kokios srities atsiliepimams klasifikuoti. Kad pagerinti rezultatus, siūloma priklausomai nuo atsiliepimų šaltinio, įtraukti daugiau atsiliepimus, jų autorius, verslą, produktus ar paslaugas apibūdinančių charakteristikų. Modelis gali būti panaudojamas ir išskiriant naudingus atsiliepimus, apie vieną dominantį produktą, paslaugą ar verslą. Atrinktuose naudinguose atsiliepimuose pateiktos vartotojų nuomonės, gali būti naudingos kuriant naują ar tobulinant jau esamą produktą, kadangi būtų atsižvelgta į vartotojų lūkesčius. Taigi naudingi atsiliepimai, padėtų pažinti verslui vartotojų poreikius.

Išvados

1. Atsiliepimai internete daro įtaką vartotojų paslaugų ir prekių pirkimo sprendimų priėmimui. Tačiau vis didėjantis atsiliepimų kiekis internetinėse svetainėse, apsunkina vartotojų naudingos informacijos, apie juos dominančią paslaugą ar produktą paiešką. Siekiama, kad vartotojui būtų sukurtos sąlygos, efektyviai atsakymų į rūpimus klausimus paieškai internetinių atsiliepimų gausoje, taip vartotojui palengvinant ir pagreitinant apsisprendimo priėmimą, dėl prekės ar paslaugos įsigijimo. Šiuo tikslu dažniausiai yra tiriamas atsiliepimų veiksmingumas, kuris apibrėžiamas vartotojų atsiliepimų internete naudingumu, bei ieškomi naudingumui įtaką darantys veiksniai. Gauti rezultatai naudojami kuriant atsiliepimų internete klasifikavimo ir rūšiavimo automatines sistemas.
2. Panaudojant teksto analitikos metodus ir R kalbą, sukurta programa, kuri importuoja vartotojų atsiliepimus, atlieka teksto gramatinį apdorojimą ir filtravimą, bei atsiliepimų sentimentų vertinimą, kurie naudojami kaip požymiai vertinant atsiliepimų naudingumą.
3. Atlikus tyrimus, nustatytas teigiamas statistiškai reikšmingas naudingumo vidurkio ryšys su *restorano vertinimu žvaigždute*, *atsiliepimų skaičiaus*, *ilgio* ir *dažnumo* kintamaisiais ($p < 0,05$). *Restorano vertinimo žvaigždute* reikšmės padidėjimas vienetu, labiausiai padidina vidutinį atsiliepimų naudingumą (1,14 karto).
4. Nustatytas teigiamas statistiškai reikšmingas ryšys tarp naudingumo vidurkio ir *atsiliepimų skaičiaus*, *ilgio* bei *dažnumo* ($p < 0,05$), o *neigiamam sentimentui* nustatytas statistiškai reikšmingas neigiamas ryšys ($p < 0,05$).
5. Geriausiai atsiliepimus klasifikavo neuroninių tinklų metodas. Teisingai klasifikuota 68,62 % nenaudingų atsiliepimų (specifiškumas), o teisingai suklasifikuotų naudingų atsiliepimų procentas (jautrumas) yra 67,81 %, nors ši klasė imtyje sudaro tik 17 %.

Literatūros sąrašas

1. SALEHAN, Mohammad ir Dan J. KIM Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics. *Decision Support Systems* [interaktyvus]. 2016, 30–40. Prieiga per doi. https://doi.org/10.1016/J.DSS.2015.10.006
2. HONG Hong, Di XU, G. Alan WANG and Weiguo FAN. Understanding the determinants of online review helpfulness: A meta-analytic investigation. *Decision Support Systems* [interaktyvus] 2017. Prieiga per doi: https://doi.org/10.1016/J.DSS.2017.06.007
3. NIELSEN: Global Advertising Consumers Trust Real Friends and Virtual strangers the Most [interaktyvus]. 2009 [žiūrėta 2018-04-05]. Prieiga per: http://www.nielsen.com/us/en/insights/news/2009/global-advertising-consumers-trust-real-friends-and-virtual-strangers-the-most.html
4. NGO-YE, Thomas L. and Atish P. SINHA. The influence of reviewer engagement characteristics on online review helpfulness: A text regression model. *Decision Support Systems* [interaktyvus]. 2014, 47-58. Prieiga per doi. https://doi.org/10.1016/j.dss.2014.01.011
5. FORMAN, Chris, Anindya GHOSE and Batia WIESENFELD. Examining the Relationship Between Reviews and Sales: The Role of Reviewer Identity Disclosure in Electronic Markets. [interaktyvu] 2008, Vol. 19, No. 3, September, pp. 291–313.Prieiga per doi: https://doi.org/10.1287/isre.1080.0193
6. EY: Big data. Changing the way businesses compete and operate. [interaktyvus]. 2014 [žiūrėta 2018-04-12]. Prieiga per: http://www.ey.com/Publication/vwLUAssets/EY_-_Big_data:_changing_the_way_businesses_operate/$FILE/EY-Insights-on-GRC-Big-data.pdf
7. WU, Jianan. Review popularity and review helpfulness: A model for user review effectiveness. Decision Support Systems [interaktyvus]. 2017, 92–103. Prieiga per doi: https://doi.org/10.1016/J.DSS.2017.03.008
8. SINGH, Jyoti Prakash and others. Predicting the "helpfulness" of online consumer reviews. Journal of Business Research [interaktyvus]. 2016, 346–355. Prieiga per doi: https://doi.org/10.1016/J.JBUSRES.2016.08.008
9. JIMENEZ, Fernando R. and Norma A. MENDOZA. Too Popular to Ignore: The Influence of Online Reviews on Purchase. *Interactive marketing* [interaktyvus]. 2013, 226-235. Prieiga per doi: https://doi.org/10.1016/j.intmar.2013.04.004
10. TIAN, Jian, Yaqi CHEN and Liwei WANG. Research on the Relationship between Online Reviews and Customer Purchase Intention: The Moderating Role of Personality Trait. *AIS Electronic Library* [interaktyvus]. 2014, [žiūrėta 2018-04-16]. Prieiga per: http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1046&context=whiceb2014

11. YIN, Dezhi, Samuel D. BOND and Han ZHANG. Anxious or angry? Efects of discrete emotions on the perceived helpfulness of online reviews. *MIS Quarterly* [interaktyvus]. 2014, Vol. 38, No. 2, pp. 539-560. Prieiga per : http://www.dennyyin.com/uploads/1/1/6/4/11646357/yin_bond_zhang_2014_misq.pdf
12. KARIMI, Sahar and Wang FANG. Online review helpfulness: Impact of reviewer profile image. *Decision Support Systems* [interaktyvus]. 2017, 39-48. Prieiga per doi: https://doi.org/10.1016/J.DSS.2017.02.001
13. ZHANG Yadong and Du ZHANG. Automatically Predicting the Helpfulness of Online Reviews. *Department of Computer Science* [interaktyvus]. 2014 [žiūrėta 2018-04-11]. Prieiga per:https://csus-dspace.calstate.edu/bitstream/handle/10211.3/121581/Report_YZhang_revised2_final.pdf?sequence=2
14. LIU, Ting, Xiao and others. Predicting movie Box-office revenues by exploiting large-scale social media content. *Multimedia Tools and Applications* [interaktyvus]. 2016, 1509–1528. Prieiga per doi: https://doi.org/10.1007/s11042-014-2270-1
15. HEAP, Bradford and others. Word Vector Enrichment of Low Frequency Words in the Bag-of-Words Model for Short Text Multi-class Classification Problems. *Cornell university library* [interaktyvus]. 2017 žiūrėta [2018-05-10]. Prieiga per: https://arxiv.org/pdf/1709.05778.pdf
16. LEE Sangjae and Joon Yeon CHOEH. Exploring the determinants of and predicting the helpfulness of online user reviews using decision trees. *Management Decision* [interaktyvus]. Vol. 55, pp.681-700 . Prieiga per doi: https://doi.org/10.1108/MD-06-2016-0398
17. DONG, Ruhai, Markus SCHAAL, Michael P. O'MAHONY, Barry SMITH. Topic Extraction from Online Reviews for Classification and Recommendation. *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence.* Dublinas, Airija. [žiūrėta 2018-05-29]. Prieiga per: https://www.ijcai.org/Proceedings/13/Papers/197.pdf
18. LEE, Pei-Ju, Yan-Han HU, Kuan-Ting LU. Assessing the helpfulness of online hotel reviews: A classification-based approach. *Telematics and Informatics* [interaktyvus]. 2018, Vol. 35, pp. 436-445. Prieiga per doi: https://doi.org/10.1016/j.tele.2018.01.001
19. REDMONK: The RedMonk Programming Language Rankings [interaktyvus]. 2017 [žiūrėta 2018-05-29]. Prieiga per: http://redmonk.com/sogrady/2017/06/08/language-rankings-6-17/
20. ČEKANAVIČIUS, Vydas ir Gediminas MURAUSKAS. *Taikomoji regresinė analizė socialiniuose tyrimuose* [interaktyvus].Vilniaus universiteto leidykla, 2014 [žiūrėta 2018-05-10]. ISBN 978-609-459-300-0. Prieiga per: http://www.statistika.mif.vu.lt/wp-content/uploads/2014/04/regresine-analize.pdf

21. BALBONIENĖ, Ingrida, Rūta BLIEKIENĖ ir Alina STUNDŽIENĖ. Ekonometrija: Praktinis regresijos ir laiko eilučių modelių taikymas: mokomoji knyga [interaktyvus]. Kaunas: Technologija, 2013 [žiūrėta 2018-05-15]. ISBN 9780210192. Prieiga per: https://www.ebooks.ktu.lt/eb/1267/ekonometrija-praktinis-regresijos-ir-laiko-eiluciu-modeliu-taikymas/
22. KRIESEL David. *Neural Networks* [interaktyvus]. Bonn, Germany, 2005 [žiūrėta 2018-05-17]. Prieiga per: http://www.dkriesel.com/_media/science/neuronalenetze-en-zeta2-1col-dkrieselcom.pdf
23. LUKOŠEVIČIŪTĖ, Kristina. Chaotinių procesų rekonstravimo bei algebrinių sekų modeliai laiko eilučių prognozavime. *Fiziniai mokslai, informatika* [interakyvus]. Kaunas, 2012 [žiūrėta 2018-05-17]. Prieiga per :https://www.personalas.ktu.lt/~kriluko/Disertacija/Disertacija_Kristinos_Lukoseviciutes.pdf
24. SMITH, D. Neural Networks: How they work, and how to train them in R. *Revolutions* [Interaktyvus]. 2017, [žiūrėta 2018-05-17]. Prieiga per: http://blog.revolutionanalytics.com/2017/03/neural-networks-r.html
25. SMITH, D. CRAN now has 10,000 R packages. Here's how to find the ones you need. *Revolutions* [Interaktyvus]. 2017, [žiūrėta 2018-05-17]. Prieiga per: http://blog.revolutionanalytics.com/2017/01/cran-10000.html
26. PAULAUSKIENĖ, Kotryna. *Duomenų tyrybos sistemų galimybių tyrimas įvairių apimčių duomenims analizuoti*. Informatikos inţinerija (09 P). Vilnius. [žiūrėta 2018-05-28] Prieiga per internetą: http://old.mii.lt/files/paulauskiene.pdf
27. BERNATAVIČIŪTĖ, Jolita. *Vizualios žinių gavybos metodologija ir jos tyrimas*. Daktaro disertacija. Technologijos mokslai, Informatikos inţinerija (07 T). Vilnius. [žiūrėta 2018-05-28] Prieiga per internetą: http://old.mii.lt/files/mii_dis_08_bernataviciene.pdf
28. ŠALNA, Bernardas ir KAMARAUSKAS Juozas. Automatinio asmens atpažinimo iš balso problemos ir perspektyvos kriminalistikoje. *Jurisprudencija* [interaktyvus]. 2005, 66(58), 142–143, [žiūrėta 2018-05-27]. Prieiga per: https://www.mruni.eu/upload/iblock/d92/019_kamarauskas.pdf

1 priedas. Duomenų paruošimas ir kintamųjų skaičiavimas

```
library(corpus)
#install.packages("dplyr")
library(dplyr)
bus <- read.csv(file="D:/JOANA/restoranu_verslas.csv")
rew   <-   read.csv(file="D:/JOANA/yelp_review.csv",nrows=400000,   header=TRUE,
sep=",")
filtered     <-     merge     (     x=rew,     y=bus,     by.x=c("business_id"),
by.y=c("business_id"),all=FALSE)
filtered <- filtered [,-c(8:9,11:12,15:17)]
rm(rew)
rm(bus)
names(filtered)[names(filtered) == 'stars.x'] <- 'zvaigzd_rest'
names(filtered)[names(filtered) == 'stars.y'] <- 'valenting'
names(filtered)[names(filtered) == 'date'] <- 'Data'
names(filtered)[names(filtered) == 'useful'] <- 'naudingumas'
names(filtered)[names(filtered) == 'review_count'] <- 'atsi_skaic'

filtered$Data <- as.Date(filtered$Data)
#Seniausias atsiliepimas
pradz_data<- "2015-01-01"
pab_data<-with(filtered, max(Data))
pab_data
filtered<-filtered[!filtered$Data < pradz_data & filtered$Data <= pab_data,]
seniausias_ats<-with(filtered, min(Data))
dim(filtered)

filtered[as.matrix(filtered)==""] <- NA
sum(is.na(filtered))

summary(filtered)
summary(filtered$naudingumas)
################################################
# ATSILIEPIMO NAUJUMAS
filtered$naujumas_ats<-pab_data-filtered$Data
filtered$naujumas_ats<-as.numeric(filtered$naujumas_ats)
#seniausias_ats<-as.Data(seniausias_ats)
#TRINTI filtered$review_old1<-filtered$Data-seniausias_ats
#filtered<-filtered[,-18]
```

```
#ATSILIEPIMO DAZNUMAS
filtered<-with(filtered, filtered[order(user_id,Data),],decreasing = TRUE)
filtered <- as_data_frame(filtered)
filtered$daznumas_ats <- sequence(rle(as.character(filtered$user_id))$lengths)

#PASKAICIUOTAS ILGIS TEKSTO (zodziu skaicius TEKSTE)
filtered$ilgis_ats <- sapply(filtered$text, function(x)
length(unlist(strsplit(as.character(x), "\\W+"))))
write.csv(filtered,"Filtered.csv",row.names = F)
```

2 priedas. Teksto valymas ir sentimentų analizė

```
library(SnowballC)
library(SentimentAnalysis)
library(textstem)
library(tm)

myCorpus <- Corpus(VectorSource(filtered$text))
myCorpus
myCorpus <- tm_map(myCorpus, removePunctuation)
myCorpus <- tm_map(myCorpus, removeNumbers)
myCorpus<- tm_map(myCorpus, tolower)
myCorpus <- tm_map(myCorpus, stripWhitespace)

removeNumPunct <- function(x) gsub("[^[:alpha:][:space:]]*", "", x)
myCorpus <- tm_map(myCorpus, content_transformer(removeNumPunct))

for (i in 1:10) {
  cat(paste0("[", i, "] "))
  writeLines(strwrap(as.character(myCorpus[[i]]), 100))
}
myStopwords <- c(stopwords(kind="english"), "xl", "two", "go", "let", "just",
"las", "vegas","le","ive")
myCorpus <- tm_map(myCorpus, removeWords, myStopwords)

for (i in 1:10) {
  cat(paste0("[", i, "] "))
  writeLines(strwrap(as.character(myCorpus[[i]]), 100))
}
myCorpus<-lemmatize_words(myCorpus)
myCorpus <- tm_map(myCorpus, stemDocument)

for (i in 10:20) {
  cat(paste0("[", i, "] "))
  writeLines(strwrap(as.character(myCorpus[[i]]), 100))
}
sentiment <- analyzeSentiment(myCorpus )
sentiment$SentimentQDAP
filtered$sentimentas<-convertToDirection(sentiment$SentimentQDAP)
```

3 priedas. Neigiama binominė regresinė analizė

```
MAT <- filtered[,-c(1:3,5:6,8:10,13:14)]

#Pirmoje eiluteje naudingumas
MAT <- MAT[, c(2, 1, 3, 4, 5,6,7,8)]

MATcopy<-MAT
MAT<-as.data.frame(MAT)

str(MAT)
#paveciamas sentimentas i skaitine reiksmes nuo -1 neigiams 0 neutralus 1
teigiamas
MAT$sentimentas<-as.integer(MAT$sentimentas)-2

max_value<-apply(MAT,2, max)
max_value
str(MAT)
########################################
library(corrplot)
correlations <-cor(MAT, y = NULL, use = "everything",
                   method =  "spearman")
correlations
corrplot(correlations, method="number", tl.col="black")

str(MAT)
MAT<-as.data.frame(MAT)
write.csv(correlations,"KORELECIJOS.csv",row.names = F)

MAT$sentimentas<-as.factor(MAT$sentimentas)
########################################
library(magrittr)
library(MASS)
summary(MAT)
#MAT$is_open<-as.numeric(MAT$is_open)
#MAT$is_open<-ifelse(MAT$is_open==1, 0,1)
#Tikriname vidurki ir dispersija Y-ko
mean(MAT$naudingumas)
var(MAT$naudingumas)
#dispersija
var(MAT$naudingumas)/mean(MAT$naudingumas)
contrasts(MAT$sentimentas) <- contr.treatment(3)
```

```
sentimentas_teig <- I(MAT$sentimentas == 1)
sentimentas_neig <- I(MAT$sentimentas == -1)
sentimentas_neig <- as.numeric(sentimentas_neig )
sentimentas_teig <-as.numeric(sentimentas_teig)

#AIC PALYGINIMUI tik su konstanta modelis
quine.nb0 <- glm.nb(naudingumas ~ 1,data = MAT)
quine.nb0
#########################################################################
summary(quine.nb2 <- glm.nb(naudingumas ~ valenting + zvaigzd_rest +
naujumas_ats + atsi_skaic
                            + ilgis_ats + daznumas_ats + sentimentas_neig
                            +sentimentas_teig,data = MAT))

anova(quine.nb2, qunine.nb0, test = "Chisq")
########################################

summary(quine.nb3 <- glm.nb(naudingumas ~ valenting + zvaigzd_rest
                            + naujumas_ats + atsi_skaic + ilgis_ats +
daznumas_ats
                            + sentimentas_neig + sentimentas_teig +
valenting*sentimentas_neig
                            + valenting*sentimentas_teig +
zvaigzd_rest*sentimentas_neig
                            +zvaigzd_rest*sentimentas_teig
+naujumas_ats*sentimentas_neig
                            +naujumas_ats*sentimentas_teig +
atsi_skaic*sentimentas_neig
                            +atsi_skaic*sentimentas_teig
+ilgis_ats*sentimentas_neig
                            +ilgis_ats*sentimentas_teig +
daznumas_ats*sentimentas_neig
                            +daznumas_ats*sentimentas_teig, data = MAT))
########################################
# MODELIS 1
Summary (quine.nb4 <- glm.nb (naudingumas ~ zvaigzd_rest
                             + atsi_skaic + ilgis_ats + daznumas_ats, data =
MAT))
Anova (quine.nb4, quine.nb0, test = "Chisq")

 (est<-cbind(ESTIMATE=coef(quine.nb4), confint(quine.nb4)))
exp(est)
```

```
########################################
# MODELIS 2
Summary (quine.nb3 <- glm.nb (naudingumas ~ + atsi_skaic + ilgis_ats +
daznumas_ats + sentimentas_neig,data = MAT))

Anova (quine.nb3, quine.nb0, test = "Chisq")

 (est<-cbind(ESTIMATE=coef(quine.nb3), confint(quine.nb3)))
exp(est)
```

4 priedas. Atsiliepimų klasifikavimas

```
install.packages("nnet")
library(tidyr)
library("nnet")
library(devtools)
library(GGally)
MAT <- filtered[,-c(1:3,5:6,8:10,13:14)]
MAT <- MAT[, c(2, 1, 3, 4, 5,6,7,8)]
MAT<-as.data.frame(MAT)
################## 1 N.B regresijos modelio kintamieji
NEURO1 <- MAT[, -c(3,5,8)]
NEURO1$naudingumas<-replace(NEURO1$naudingumas, NEURO1$naudingumas < 2,0)
NEURO1$naudingumas<-replace(NEURO1$naudingumas, NEURO1$naudingumas > 1,1)
NEURO1$naudingumas<-as.factor(NEURO1$naudingumas)
NEURO1<-as.data.frame(NEURO1)
write.csv(NEURO1,file="NEURO1.csv")

data<-NEURO1
####### klasifikavimas ########
library(devtools)
install_github("davidavdav/ROC")
library(ROC) # package for ROC, DET, AUC, EER
library(precrec) # Precision-Recall curves
library(tidyr) # spread function (unpivot)
library(ggplot2)

library(caret) # center/scale and k-fold CV
library(randomForest)
#library(MASS) # LDA
library(FNN) # k-NN

# neural network packages
#library(RSNNS)
#library(neuralnet)
library(nnet)
library(corrplot)
rootFolder <- dirname(sys.frame(1)$ofile)
setwd(rootFolder)

factorsNumeric <- function(d) modifyList(d, lapply(d[, sapply(d, is.factor)],
as.numeric))
```

```
D <- read.csv("C:/Users/Joana/Documents/NEURO2.csv",sep=";",header=T)

myData <-D

colY <- "naudingumas"
idxY <- which(colnames(myData) %in% colY)
myData[,idxY] <- factor(myData[,idxY],labels=c("False","True"))

myDataCORR<-cor(myData[,-idxY])
corrplot(myDataCORR, method = "number")

procValues <- preProcess(factorsNumeric(myData[,-idxY]), method = c("center",
"scale"))

# random forest settings
mtry <- floor(sqrt(ncol(myData)-1))
ntree <-500
ptree <- 0

# k-NN
knn_neighs <- 99
hiddenSize <- 10

k <- 3 #   duomeni suskaido i tris intervalus
myFolds <- createFolds(myData[,colY],k)

rm(D)
gc()
myResults <- NULL

for (i in 1:k) {

  tstInd <- myFolds[[i]]
  trnIdx <- as.logical(rep(1,1,nrow(myData)))
  trnIdx[tstInd] <- FALSE
  trnInd <- which(trnIdx)
  target <- as.logical(myData[tstInd,idxY])

  nmd <- names(myData) # long for neuralnet, short for logit and LDA
  formulaLong <- as.formula(paste(paste(colY," ~",sep=""), paste(nmd[!nmd %in%
colY], collapse = " + ")))
  formulaShort <- as.formula(paste(paste(colY,".",sep="~")))
```

```
  cat(sprintf("CV fold %d out of %d / Random Forest\n", i, k))
  model_classwt <- prop.table(table(myData[trnInd,idxY]))

  rf_model <- tuneRF(myData[trnInd,-idxY], myData[trnInd,idxY], mtryStart =
mtry, ntreeTry = ntree,
                     stepFactor = 2, improve = 0.01, plot=FALSE, doBest=TRUE,
                     classwt = model_classwt, cutoff = model_classwt,
                     strata = Y, replace = FALSE, importance=FALSE, do.trace =
ptree)
  model <- rep("RF",length(target))
  soft <- predict(rf_model,myData[tstInd,-idxY],type="prob")
  score <- soft[,2]
  myResults <- rbind(myResults,data.frame(tstInd,model,score,target))
  rm(rf_model)

  cat(sprintf("CV fold %d out of %d / k-Nearest Neighbors\n", i, k))
  knn_model <- knn(predict(procValues, factorsNumeric(myData[trnInd,-idxY])),
predict(procValues, factorsNumeric(myData[tstInd,-idxY])), myData[trnInd,idxY],
k = knn_neighs, prob = TRUE, algorithm = "kd_tree")
  model <- rep("kNN",length(target))
  score <- 1-abs(as.numeric(knn_model)-1-attr(knn_model,"prob"))
  myResults <- rbind(myResults,data.frame(tstInd,model,score,target))
  rm(knn_model)

  cat(sprintf("CV fold %d out of %d / Neural Network (nnet package)\n", i, k))
  nnet_model <- nnet(predict(procValues, factorsNumeric(myData[trnInd,-idxY])),
as.numeric(myData[trnInd,idxY])-1, size = hiddenSize, entropy = TRUE, maxit =
1000, trace=FALSE)
  model <- rep("nnet", length(target))
  score <- predict(nnet_model, predict(procValues,
factorsNumeric(myData[tstInd,-idxY])), type="raw")
  myResults <- rbind(myResults,data.frame(tstInd, model,score,target))
  rm(nnet_model)
}

myModels <- levels(myResults[,"model"])
myScores <- spread(myResults, model, score)
write.csv(myScores,file="myResults.csv")

# ROC curves
myModelNames <- NULL
i <- 1
```

```
performance                                                                    <-
roc.plot(myResults[myResults[,"model"]==myModels[i],],i,traditional=TRUE)
myModelNames[i] <- sprintf('%s AUC=%5.3f',myModels[i],1-performance['pAUC'])
for (i in 2:length(myModels)) {
  performance                                                                  <-
roc.plot(myResults[myResults[,"model"]==myModels[i],],i,traditional=TRUE)
  myModelNames[i] <- sprintf('%s AUC=%5.3f',myModels[i],1-performance['pAUC'])
}
legend(0.3,0.5,myModelNames,lty=rep(1,1,length(myModels)),col=1:length(myModels)
)
# DET curves
myModelNames <- NULL
det.plot(NULL,1,xmax=75,ymax=75)
for (i in 1:length(myModels)) {
  performance <- det.plot(myResults[myResults[,"model"]==myModels[i],],nr=i+1)
  myModelNames[i] <- sprintf('%s EER=%5.2f%%',myModels[i],performance['eer'])
}
legend(-3.2,-
1.24,myModelNames,lty=rep(1,1,length(myModels)),col=2:(length(myModels)+1))

# Precision-Recall curves
myScores <- spread(myResults, model, score)
msmdat <- mmdata(myScores[,-c(1,2)], myScores[,2], posclass = TRUE, modnames =
myModels)
plot(autoplot(evalmod(msmdat), "PRC", type="b"))

myScores$target <- as.numeric(myScores$target)
write.csv(myScores,file="churn_scores.csv")

require(OptimalCutpoints)
oc <- optimal.cutpoints(X = "nnet", status = "target", tag.healthy = 1, methods
= "SpEqualSe", data = myScores)
#Slenkscio reiksme
oc
#SUMAISYMU MATRICA
library(e1071)
caret::confusionMatrix(myScores$nnet>oc,as.logical(myScores$target),mode="everyt
hing")
plot(nnet)
```