VILNIAUS UNIVERSITETAS

TOMAS PRANCKEVIČIUS

DAUGIAKLASIŲ TEKSTINIŲ DUOMENŲ KLASIFIKAVIMO METODŲ TYRIMAS

Daktaro disertacijos santrauka

Technologijos mokslai, informatikos inžinerija (07 T)

Vilnius, 2017

Disertacija rengta 2011–2016 metais Vilniaus universitete, Matematikos ir informatikos institute.

Mokslinis vadovas:

dr. Virginijus Marcinkevičius (Vilniaus universitetas, technologijos mokslai, informatikos inžinerija – 07 T).

Disertacija ginama viešame Gynimo tarybos posėdyje:

Pirmininkas:
prof. dr. Rimantas Vaicekauskas (Vilniaus universitetas, technologijos mokslai, informatikos inžinerija – 07 T).

Nariai:
prof. dr. Tomas Krilavičius (Vytauto Didžiojo universitetas, technologijos mokslai, informatikos inžinerija – 07 T),
prof. dr. Audrius Lopata (Vilniaus universitetas, technologijos mokslai, informatikos inžinerija – 07 T),
prof. dr. Audris Mockus (Tenesio universitetas, JAV, technologijos mokslai, informatikos inžinerija – 07 T)
prof. dr. Darius Plikynas (Vilniaus universitetas, technologijos mokslai, informatikos inžinerija – 07 T).

Disertacija bus ginama Vilniaus universiteto viešame Gynimo tarybos posėdyje 2017 m. rugsėjo 28 d. 11 val. 00 min. Vilniaus universitete 203 auditorijoje. Adresas: Akademijos g. 4, LT-08663 Vilnius, Lietuva.

Disertacijos santrauka išsiųsta 2017 m. rugpjūčio 26 d.

Disertaciją galima peržiūrėti Vilniaus universiteto bibliotekoje ir VU interneto svetainėje adresu: www.vu.lt/lt/naujienos/ivykiu-kalendorius.

VILNIUS UNIVERSITY

TOMAS PRANCKEVIČIUS

INVESTIGATION OF MULTI-CLASS CLASSIFICATION METHODS FOR TEXTUAL DATA

Summary of Doctoral Dissertation

Technological Sciences, Informatics Engineering (07 T)

Vilnius, 2017

The dissertation was written from 2011 to 2016 at Vilnius University Institute of Mathematics and Informatics.

Scientific Supervisor:

Dr. Virginijus Marcinkevičius (Vilnius University, Technological Sciences, Informatics Engineering – 07 T).

The thesis will be defended at the Council at Vilnius University:

Chairman:
Prof. Dr. Rimantas Vaicekauskas (Vilnius University, Technological Sciences, Informatics Engineering – 07 T).

Members:
Prof. Dr. Tomas Krilavičius (Vytautas Magnus University, Technological Sciences, Informatics Engineering – 07 T),
Prof. Dr. Audrius Lopata (Vilnius University, Technological Sciences, Informatics Engineering – 07 T),
Prof. Dr. Audris Mockus (University of Tennessee, USA, Technological Sciences, Informatics Engineering – 07 T),
Prof. Dr. Darius Plikynas (Vilnius University, Technological Sciences, Informatics Engineering – 07 T).

The thesis will be defended at the public session of the Scientific Council in the auditorium number 203 at Vilnius University at 11:00 a. m. on the 28th of September 2017.

The summary of the doctoral thesis was distributed on the 26th of August 2017.

A copy of the doctoral thesis is available for review at the Library of Vilnius University or on this website: www.vu.lt/lt/naujienos/ivykiu-kalendorius.

1. ĮVADAS

Šiuolaikiniai kompiuteriniai skaičiavimo resursai didina galimybes atlikti įvairias duomenims imlias natūralios kalbos apdorojimo ir sistemos mokymosi užduotis. Viena iš tokių užduočių – daugiaklasių tekstinių duomenų klasifikavimas į iš anksto nustatytas klases. Teksto klasifikavimo problematiką naudojant didelius duomenis (angl. *Big data*) pastaruoju metu plačiai tiria daugelis mokslininkų. Teksto klasifikavimas dažnai naudojamas kaip vienas iš sentimentų analizės būdų, leidžiantis analizuoti tam tikro turinio emocinį toną ir klasifikuojant priskirti tekstui tam tikras sentimentų reikšmes – teigiamą arba neigiamą. Teksto klasifikavimas apima įvairias sritis, suteikiančias technines galimybes klasifikuoti didelės apimties tekstinius duomenis, tarp jų matematinius, statistinius, duomenų inžinerijos, mokymosi, modeliavimo, didelės spartos skaičiavimų ir natūralios kalbos apdorojimo metodus bei būdus. Kitaip tariant, tai naujos duomenų analizės formos, paremtos žinių rinkimu, naudojant atminčiai imlius kompiuterius ir jų tinklus.

Teksto klasifikavimas yra sudedamoji dirbtinio intelekto dalis ir nauja besivystanti sritis, apimanti natūralios kalbos duomenų rinkimą ir analizę bei pateikianti galimas alternatyvas sprendimus priimantiems asmenims. Šio tyrimo tikslas yra palyginti daugiaklasius klasifikavimo metodus, įvertinant jų teksto klasifikavimo tikslumą, priklausomybę nuo mokymų duomenų dydžio, *n*-gramų skaičiaus, žodžių leksemų (angl. *word tokens*) kalbos dalių, žodžių dažnių ir kt. Eksperimentinėje dalyje analizuojami atsiliepimų apie prekes duomenys, gauti iš Amazon[1] prekyvietės. Atsiliepimų apie prekes duomenys kaupiami todėl, kad jie gali būti naudingi kiekvienam potencialiam pirkėjui padėti apsispręsti, ar užsisakyti tam tikras prekes ar paslaugas. Kita vertus, kai kurie atsiliepimai apie prekes yra nenaudingi, t. y. juose nepateikiama svarbios informacijos ar pateikiamas neigiamas vertinimas, pavyzdžiui, dėl to, kad prekės buvo pristatytos diena vėliau, nei tikėtasi. Be to, esama ir tokių atsiliepimų apie prekes, kuriuose nėra jokios informacijos, nei naudingos, nei nenaudingos.

Ištirti tekstinius duomenis yra sudėtingas uždavinys, todėl informatikos inžinerija stengiasi sukurti naujoviškų įrankių ir metodų, kurie leistų efektyviai pasiekti šį tikslą. Šioje disertacijoje tiriama debesijos kompiuterijos technologijų įtaka duomenų klasifikavimui ir analizuojami šiuolaikiniai natūralios kalbos apdorojimo metodai. Tyrimai ir eksperimentai vykdyti naudojant Apache Spark[2] programinę įrangą, pritaikytą ir veikiančią atminčiai imliose kompiuterinėse platformose, tokiose kaip debesų kompiuterija. Šios disertacijos tikslas – pasiūlyti duomenų požymių ir klasifikatorių derinį, kuris leistų pasiekti didesnį daugiaklasių klasifikavimo metodų klasifikavimo tikslumą, kai klasifikuojami didelės apimties trumpų tekstinių atsiliepimų apie prekes duomenys.

Problemos formulavimas

Tekstiniai duomenys yra vertingas informacijos šaltinis, kuriuo gaunamos ir analizuojamos žinios apie vartotojų internetinę ir pirkimo elgseną bei emocijų nulemtą sprendimą, ką pirkti. Atsiliepimai apie prekes paprastai turi didelės įtakos vartotojo emocijoms ir sprendimui pirkti ar nepirkti tam tikros prekės ar paslaugos. Internete

[1] Amazon yra registruotas prekės ženklas. Plačiau: https://amazon.com
[2] Apache Spark yra registruotas prekės ženklas. Plačiau: https://www.apache.org

prekiaujantys mažmenininkai visada kreipia dėmesį į priemones ir metodus, skatinančius produktų pirkimą, didinančius pardavimų pajamas ir padedančius gauti iš vartotojų grįžtamąjį ryšį. Produktų apžvalgos ar atsiliepimai (angl. *product-reviews*), tai sudėtingi ir didelės apimties tekstai. Vartotojai ir pardavėjai šio tipo tekstus naudoja perkant ar parduodant prekes internete, t. y. specializuotose internetinėse prekyvietėse. Tačiau kai kurie atsiliepimai apie prekes ar paslaugas yra naudingesni ir turi didesnės įtakos vartotojams bei pardavėjams nei kiti, nors juos išanalizuoti yra sudėtinga dėl didelio jų kiekio. Atsiliepimai apie prekes gali būti profesionalūs, neprofesionalūs, trumpi, ilgi, apimantys tam tikrus psichologinius, elgsenos, suvokimo ir vertinimo aspektus, priklausančius nuo skirtingų asmenybės tipų. Mokslinėje literatūroje yra keletas darbų, kuriuose analizuojama atsiliepimų apie prekes klasifikavimo problema, tačiau šių darbų autoriai nesiūlo daugiaklasių tekstinių duomenų klasifikavimo metodų, kurie padėtų užtikrinti didesnį klasifikavimo tikslumą apdorojant didelės apimties tekstinius duomenis, kai tam naudojami duomenų požymių parinkimo būdai apjungiant *n*-gramas. Lyginamuosiuose tyrimuose paprastai analizuojami Bajeso (angl. *Naïve Bayes*) ir atraminių vektorių mašinų (angl. *Support vector machine*) metodai, tačiau neanalizuojamas logistinės regresijos (angl. *Logistic Regression)* metodas kaip vienas iš galimų klasifikavimo metodų. Be to, kai kurie autoriai yra įrodę, kad klasifikavimo metodų parametrų keitimas turi mažiau įtakos klasifikavimo tikslumui nei tekstyno paruošimas pritaikius natūralios kalbos apdorojimo ir duomenų požymių parinkimo metodus.

Atsižvelgiant į tai, šios disertacijos tikslas – sukurti duomenų požymių parinkimo ir klasifikatorių derinį, kuris padėtų išspręsti atsiliepimų apie prekes klasifikavimo problemą, siekiant pagerinti klasifikavimo tikslumą, kai klasifikuojami didelės apimties atsiliepimų apie prekes duomenys. Kitaip tariant, pagrindinis disertacijos siekis yra nustatyti ir tiksliai priskirti tam tikrai klasei nežinomus atsiliepimus apie prekes, kai turima atsiliepimų apie prekes mokymo aibė su pažymėtomis klasėmis. Šiuolaikinių atminčiai imlių kompiuterinių technologijų evoliucija šiandien suteikia naujas galimybes naujoviškiems didelės apimties natūralios kalbos duomenų apdorojimo ir klasifikavimo būdams. Šis mokslinis tyrimas yra apie didelės apimties natūralios kalbos duomenų pokyčių poveikį, duomenų analitikos karkasų vystymąsi, sistemos mokymąsi, ir daugiaklasius klasifikavimo metodus.

Tyrimo objektas

Tyrimo objektas – daugiaklasiam klasifikavimui skirti natūralios kalbos apdorojimo metodai, paremti moderniais debesų kompiuterijos technologiniais sprendimais ir duomenų analitikos karkasais.

Tyrimo tikslas ir uždaviniai

Tyrimo tikslas – sukurti duomenų požymių parinkimo ir klasifikatorių derinį, kuris padėtų išspręsti atsiliepimų apie prekes klasifikavimo problemą, siekiant pagerinti klasifikavimo tikslumą, kai klasifikuojami didelės apimties atsiliepimų apie prekes duomenys. Tyrimo tikslui įgyvendinti iškelti šie uždaviniai:

1. Ištirti technologijas dirbančias su dideliais duomenims, daugiaklasius klasifikavimo ir natūralios kalbos apdorojimo metodus.

2. Palyginti technologijas dirbančias su dideliais duomenims, daugiaklasius klasifikavimo metodus ir duomenų požymių parinkimo metodus, taikomus didelės apimties tekstiniams duomenims, lyginamosios analizės būdu palyginti realizuotus ir ištirtus metodus naudojant realias duomenų aibes, naudojant šiuos daugiaklasių duomenų klasifikavimo kriterijus: klasifikavimo tikslumas, preciziškumas, prisiminimas, klaidų lygis ir F1 matavimas.
3. Pasiūlyti duomenų požymių parinkimo būdą, kuris leistų padidinti klasifikavimo tikslumą, kai klasifikuojami trumpi atsiliepimų apie prekes žinučių tekstai.
4. Pasiūlyti modifikuotą darbų sekos (angl. *workflow*) modelį, kuris apimtų technikas ir daugiaklasius klasifikavimo metodus, tinkamus tiksliau klasifikuoti trumpąsias teksto žinutes.

Tyrimo metodai

Tyrimo metodologija, taikoma šioje disertacijoje, pagrįsta šiais metodais:

1. Bibliografiniu suformuluotų klausimų ir uždavinių tyrimu, skirtu nustatyti, atrinkti ir įvertinti tyrimo įrodymus, tiesiogiai susijusius su analizuojamais klausimais ir uždaviniais.
2. Mokslinių, eksperimentinių ir praktinių pasiekimų analizė šiose srityse, kaip sistemos mokymasis klasifikuoti tekstinius duomenis debesijos kompiuterinėse technologijose ir informacijos išgavimo, organizavimo, analizavimo, lyginimo bei kaupimo metodų naudojimas.
3. Kiekybinės ir kokybinės informacijos rinkimu, kuris buvo pasitelktas problemų sprendimo procedūroms, pvz., rinkti eksperimentiniams duomenims apie daugiaklasius klasifikavimo metodus.
4. Konstruktyvaus tyrimo procedūra, kuri leido suformuoti naujus darinius, galinčius pasiūlyti sprendimus realaus pasaulio iššūkiams bei prisidėti prie srities, kurioje šie dariniai gali būti taikomi, teorijos plėtros.
5. Atvejais pagrįstų (angl. *case base*) ir kontroliuojamų eksperimentų metodų, naudojamu eksperimentinėje šios disertacijos dalyje.
6. Programinės įrangos plėtros metodais, kurie buvo naudojami daugiaklasių klasifikavimo metodų kūrimui eksperimentų etape ir duomenų požymių parinkimo metodų, taikytinų kurti didelės apimties tekstiniams duomenims pagal daugiaklasių klasifikavimo kriterijus – klasifikavimo tikslumą, preciziškumą, prisiminimą, klaidų lygį ir F1 rodiklį.

Tyrimo mokslinė vertė

Tyrimas yra moksliškai vertingas dėl šių priežasčių:

1. Disertacijoje eksperimentiškai ištirti didelės apimties tekstinių duomenų – trumpųjų atsiliepimų apie prekes – daugiaklasiai klasifikavimo metodai.
2. Siūlomi *n*-gramų junginiai (unigramos, bigramos ir trigramos) yra efektyvi žodžių dažnių duomenų požymių atrankai, naudojant logistinės regresijos (angl. *Logistic Regression)*, atraminių vektorių mašinų (angl. *Support Vector Machine*), Bajeso (angl. *Naïve Bayes*), atsitiktinio miško (angl. *Random Forest*), sprendimų medžio (angl. *Decision Tree*) ir daugiasluoksnio perceptrono (*angl. Multilayer Perceptron*)

klasifikatorius. Šie junginiai leidžia išspręsti daugiaklasių klasifikavimo problemą ir pasiekti didesnį klasifikavimo tikslumą.

3. Disertacijoje pateikiama daugiaklasių tekstinių duomenų klasifikavimo metodika, derinama su duomenų požymių parinkimo metodais (triukšmo mažinimo, žodžių krepšelių ir žodžių dažnių), yra tinkama klasifikuoti duomenų rinkiniams, apimantiems didelį kiekį trumpų tekstų, tokių kaip atsiliepimai apie prekes. Šio tipo duomenys dažniausiai naudojami pirkti ar parduoti prekėms internetu, pvz., specializuotose internetinėse duomenų saugyklose ar specializuotuose interneto kataloguose.
4. Lyginamoji didelės apimties duomenų analitikos karkasų analizė parodė, kad Apache Spark analitikos karkasas, naudojama debesų kompiuterijos technologijose, yra tinkama klasifikuoti didelės apimties tekstiniams duomenims ir gali būti taikoma įvairiems klasifikavimo algoritmams realizuoti horizontalaus paskirstymo kompiuterių tinkluose.

Tyrimo praktinė vertė

Siūlomi daugiaklasių tekstinių duomenų klasifikavimo bei duomenų požymių parinkimo metodai gali būti naudojami įvairiose didelės apimties tekstinių duomenų apdorojimo sistemose ir įrankiuose:

1. Dirbant su didelės apimties duomenimis, šie metodai gali padėti atrinkti ir suklasifikuoti neigiamus, teigiamus ir neutralius atsiliepimus apie prekes ir atrinkti tik tiksliausius teigiamus atsiliepimus, kurie leistų padidinti įvairių prekių ar paslaugų pardavimo pajamas, teikti papildomas pardavimų paslaugas ar vykdyti tam tikras vartotojų išlaikymo programas, pvz., nustatyti nepatenkintus vartotojus.
2. Siūlomi metodai gali padėti rasti optimalias duomenų struktūras ir nustatyti jų vertę, bei realizuoti algoritmus. Jie taip pat gali padėti suprasti ir prognozuoti tekstinius duomenis, kai priimami sprendimai ir renkamos žinios mokslo bei verslo srityse.
3. Siūlomi metodai gali padėti analizuoti didelius tinklinių duomenų rinkinius rinkodaros tyrimuose, internete aptinkant antisocialinės elgsenos požymius ar klasifikuojant tekstinius dokumentus, susijusius su kibernetine sauga, pvz., su kenkėjiškais srities vardais (angl. *domains*), programų kodo pažeidžiamumu, sukčiavimo atvejų nustatymu ir pan.

Ginamieji teiginiai

Disertacijos ginamieji teiginiai yra šie:

1. Didelės apimties duomenims skirti analitikos karkasai gali būti sėkmingai naudojamos atminčiai imlioms operacijoms, skirtoms sistemos mokymui, pvz., klasifikavimo algoritmuose.
2. Siūlomas duomenų požymių parinkimo metodas, pagrįstas *n*-gramų (unigramos, bigramos ir trigramos) ir žodžių dažnių metodų junginiais, leidžia pasiekti didesnį klasifikavimo tikslumą klasifikuojant trumpus tekstus, tokius kaip atsiliepimų apie prekes žinutes, kai šis metodas naudojamas kartu su logistinės regresijos (angl. *Logistic Regression)*, atraminių vektorių mašinų (angl. *Support Vector Machine*), Bajeso (angl. *Naïve Bayes*), atsitiktinio miško (angl. *Random Forest*), sprendimų

medžio (angl. *Decision Tree*) ir daugiasluoksnio perceptrono (*angl. Multilayer Perceptron*) metodais.

3. Logistinės regresijos metodas pasiekia geresnius rezultatus nei atraminių vektorių mašinų, Bajeso, atsitiktinio miško, sprendimų medžio ir daugiasluoksnio perceptrono klasifikavimo metodai, kai klasifikuojami didelės apimties daugiaklasiai tekstinių duomenų rinkiniai, kuriems naudojami siūlomi *n*-gramų (unigramos, bigramos ir trigramos) junginiai ir lyginami su unigramos atveju, naudojant žodžių dažnių bei triukšmo mažinimo technikas, tokias kaip žodžių segmentavimas, nereikšmingų žodžių šalinimas, didžiųjų raidžių keitimas mažosiomis ir žodžio grąžinimas į pagrindinę formą.
4. Siūlomas daugiaklasių tekstinių duomenų klasifikavimo metodas kartu su *n*-gramų junginiais (unigramos, bigramos ir trigramos) leidžia pasiekti didesnį klasifikavimo tikslumą klasifikuojant trumpus tekstus, tokius kaip atsiliepimai apie prekes.

Tyrimo aprobavimas ir publikavimas

Pagrindiniai šios disertacijos rezultatai buvo paskelbti šiose mokslinėse publikacijose:

Straipsniuose recenzuojamuose periodiniuose moksliniuose žurnaluose:

1. Pranckevičius T. Parallel data processing services based on Cloud computing technology. *Information Sciences,* 73: 64–73, 2015. ISSN 1392-0561.

2. Pranckevičius T., Marcinkevičius V. Comparison of Naïve Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification. *Baltic J. Modern Computing*, Vol. 5, No. 2, 221–232, 2017. ISSN 2255-8950.

Straipsniuose recenzuojamuose konferencijų leidiniuose:

3. Pranckevičius T., Marcinkevičius V. Logistic Regression and Tokenization Methods Applied for Multi-Class Text Classification. *2016 IEEE 4th Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE)*, Vilnius, 2016. ISBN: 978-1-5090-4473-3.

Straipsniai kitose konferencijų leidiniuose:

4. Pranckevičius T., Marcinkevičius V. Classification and visualization algorithms on cloud computing: issues and advantages. *Data analysis methods for software systems: 7th international workshop* [abstract book], Druskininkai, December 3–5, Vilnius: Vilniaus universiteto Matematikos ir informatikos institutas, 2015. ISBN 978-9986-680-58-1.

Tarptautinių mokslinių konferencijų pristatymuose:

5. Pranckevičius T. Comparison of Naive Bayes and Random Forest classifiers on product-review data. *28th European Conference on Operational Research.* Poznan, Poland. July 3–6, 2016.
6. Pranckevičius T. Cloud computing and applications based on software services. *2nd Workshop on Software Services.* Timisoara, Romania. November 11–14, 2011.

Tarptautinių mokslinių konferencijų Lietuvoje pristatymuose:

7. Pranckevičius T. Application of Logistic Regression with Part-Of-The-Speech Tagging for Multi-Class Text Classification. *The 4th Workshop on AIEEE'16, Vilnius, Lithuania.* November 10–12, 2016.

8. Pranckevičius T. Parallel data processing services based on cloud computing technology. *56th International conference: Computer Days – 2015*. Panevėžys, Lithuania. September 17–19, 2015.

9. Pranckevičius T. Classification and visualization algorithms on cloud computing: issues and advantages. *7th International Workshop: Data Analysis Methods for Software Systems*. Druskininkai, Lithuania. December 3–5, 2015.

10. Pranckevičius T. Investigation of impact of cloud computing technology on the visualization and classification algorithms. *International doctoral consortium. Informatics and Informatics Engineering Education Research: Methodologies, Strategies and Implementation*. Druskininkai, Lithuania. November 30–December 4, 2011.

2. TYRIMO METODIKA

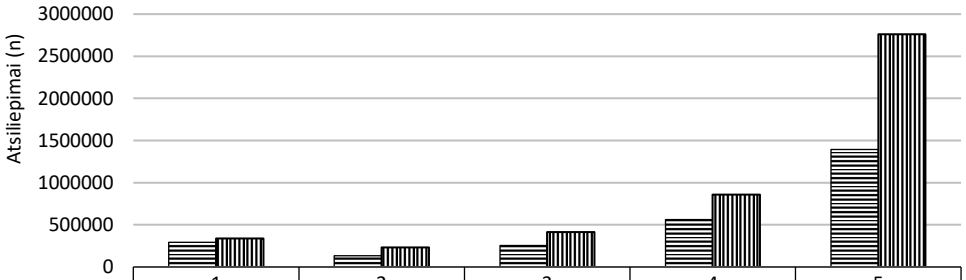
Pasiūlytas modifikuotas darbų sekos modelis (3 pav.), apimantis atitinkamus natūralios kalbos apdorojimo metodus daugiaklasių tekstinių duomenų klasifikavimui, leidžia palyginti ir įvertinti užduoties atlikimo kriterijus, t. y. išmatuoti klasifikavimo tikslumą. Šis modifikuotas darbų sekos modelis buvo naudotas klasifikuoti trumposioms tekstinėms žinutėms, šiuo modeliu taip pat remtasi atliekant eksperimentus.

Informacija apie duomenų rinkinius

Siekiant pagrįsti eksperimentų rezultatus ir mokslines išvadas, tyrimui pasirinkti du atvirai prieinami duomenų rinkiniai: Amazon prekyvietės klientų atsiliepimai apie Android programėles (duomenų rinkinys A) ir Amazon klientų atsiliepimai apie filmus ir TV (televiziją) (duomenų rinkinys B). Bendrasis Amazon klientų atsiliepimų apie Android programėles kiekis yra išreiškiamas $n = 2638274$. Bendrasis Amazon klientų atsiliepimų apie filmus ir TV kiekis yra išreiškiamas $n = 4607047$. Abiejose išraiškose nėra pasikartojančių duomenų. Žemiau pateikiamas atsiliepimo pavyzdys:

```
["reviewerID": "AUI0OLXAB3KKT", "asin": "B004A9SDD8", "reviewerName": "A
Customer", "helpful": [0, 0], "reviewText": "Glad to finally see this app on
the android market. My wife has it on her iPhone and iPad and my son (15 months)
loves it! Hopefully more apps like this are on the way!", "overall": 5.0,
"summary": "Great app!!!", "unixReviewTime": 1301184000, "reviewTime": "03 27,
2011"],
```

čia: *reviewerID* – naudotojo identifikacijos duomenys; *asin* – identifikacijos numeris; *reviewerName* – naudotojo vardas; *helpful* – naudotojų, kuriems atsiliepimas padėjo, pateiktas grįžtamasis ryšys apie atsiliepimo kokybę ir naudingumą; *reviewText* – kliento atsiliepimas apie produktą; *overall* – kliento pateikiamas produkto vertinimas (šiame tyrime vertinimai skirstomi balais nuo 1 iki 5: 1 yra žemiausias, o 5 – aukščiausias galimas vertinimas); *summary* – kliento atsiliepimo santrauka; *(unix)ReviewTime* – (*Unix*) atsiliepimo laikas. Eksperimento metu naudoti tik *overall* ir *reviewText* (atsiliepimo tekstas) duomenų laukeliai. Duomenis sudaro įvairūs klientų atsiliepimai, išreiškiami $D = \{d_1, d_2, d_3 \dots d_n\}$, kai *n* yra bendrasis atsiliepimų skaičius.



	1	2	3	4	5
Atsiliepimų klasėje (A)	294293	133904	253586	561829	1394662
Atsiliepimų klasėje (B)	339544	233221	415369	857505	2761408

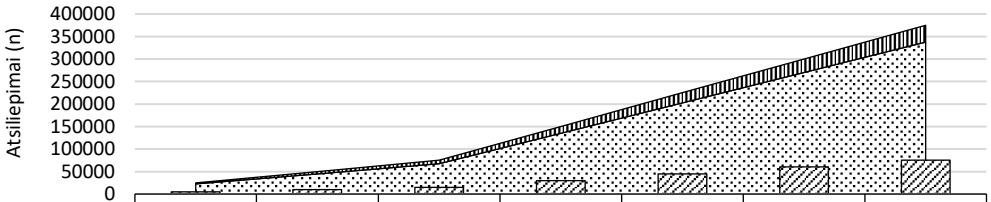
Klasė (C_i)

1 pav. Klientų atsiliepimų apie prekes pasiskirstymas pagal klases

Šie atsiliepimai klasifikuojami pagal skirtingas klases, t. y. tam tikra kategorija priskiriama tam tikram atsiliepimui pagal vertinimo skaitinę reikšmę $C = \{c_1, c_2, c_i \dots c_5\}$,

kai C_i ($C_i = i$, ir i reiškia klasės indeksą), o m yra bendrasis klasių skaičius ($m = 5$), ir ši kategorija yra laikoma klase. Duomenų klasės paskirstymas C_i duomenų rinkinyje pavaizduotas 1 pav.

Siekiant pagerinti klasifikavimą, nuspręsta padalyti duomenis į tolygiai pasiskirsčiusius duomenų rinkinius kiekvienai klasei naudojant asimetrišką (angl. *skewness*) duomenų išsidėstymo metodą, kad kiekvienoje klasėje būtų surinktas vienodas skaičius atsiliepimų apie prekes įrašų.



	DS1	DS2	DS3	DS4	DS5	DS6	DS7
Testavimas	2500	5000	7500	15000	22500	30000	37500
Apmokymas	22500	45000	67500	135000	202500	270000	337500
Atsiliepimų klasėje	5000	10000	15000	30000	45000	60000	75000

2 pav. Duomenų struktūra apmokymui ir testavimui

Eksperimentuose naudoti septyni vienodai pasiskirstę įvairių dydžių duomenų rinkiniai DS1, DS2, DS3, DS4, DS5, DS6 ir DS7. Duomenų rinkinių struktūra apmokymui ir testavimui pasiskirsčiusi taip: 90 % duomenų skirta apmokymui, o 10 % – testavimui, esant vienodam skaičiui atsiliepimų apie prekes kiekvienoje klasėje. Duomenų struktūra apmokymui ir testavimui iliustruojama 2 pav.

Duomenų analitikos karkasas teksto klasifikavimui

Šiame tyrime eksperimentams naudotasi duomenų apdorojimo klasterio infrastruktūra. Duomenų klasifikavimo užduotys atliktos naudojant MLlib biblioteką Apache Spark kompiuterinėje platformoje su šia konfigūracija:

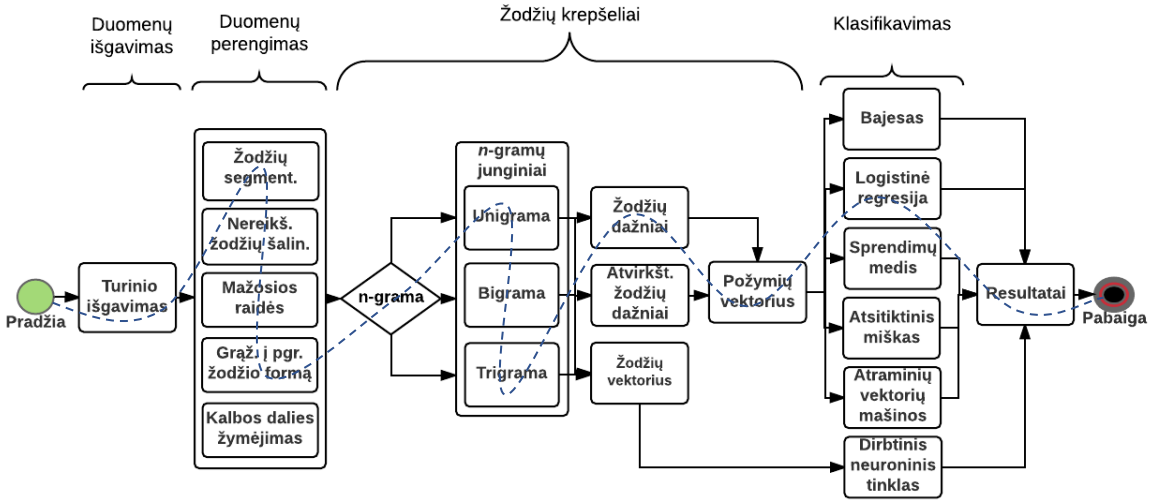
- Infrastruktūra įdiegta Google[3] debesijos platformoje. Visi eksperimentai vykdyti naudojant Apache Spark v1.6.2, Python v2.7.6 ir NLTK v3.0.0.
- Naudotas Apache Spark duomenų apdorojimo klasteris, apimantis ir sistemos mokymosi įrankį MLlib, kuris buvo įdiegtas Google debesijos platformoje. Pagrindinį mazgą sudaro 2 vCPU (virtualus centrinis procesorius) ir 26 GB operatyvios atminties, o kitus du mazgus sudaro 2 vCPU su 13 GB operatyvios atminties kiekvienam mazgui.

Darbų sekos modelis

Atsiliepimų apie prekes duomenims apdoroti sukurtas darbų sekos modelis (žr. 3 pav.) siekiant palyginti Bajeso, atsitiktinio miško, sprendimų medžio, atraminių vektorių mašinų, logistinės regresijos ir daugiasluoksnio perceptrono metodus. Atsiliepimams apie prekes apdoroti darbų sekos modelis naudotas šiame tyrime, o 3 pav. pažymėtas takelis atspindi geriausią klasifikavimo metodą, kuris naudojamas kartu su duomenų požymių

[3] Google yra registruotas prekės ženklas. Plačiau: https://www.google.com

parinkimo metodais. Toliau šiame poskyryje atsiliepimams apie prekes apdoroti skirtas darbų sekos modelis aprašomas plačiau. Darbų sekos modelį sudaro keturi pagrindiniai etapai: duomenų išgavimas, atsiliepimų apie prekes teksto parengimas (triukšmo mažinimas), žodžių krepšeliai ir teksto transformavimas į požymių vektorių ir klasifikavimo algoritmų pritaikymas. Pirma, pagrindinis duomenų gavimo etapo tikslas yra iš duomenų išrinkti tik reikiamus ir susijusius duomenis. Antra, duomenys paruošiami sistemos mokymosi etapui. Šiame etape duomenys konvertuojami, t. y. transformuojami ir jiems pritaikomi duomenų požymių parinkimo kriterijai. Pagrindinis šio etapo tikslas yra atrinkti svarbiausius duomenų laukelius, reikalingus tyrimui. Trečia, pagrindinis sistemos mokymosi etapo tikslas yra vykdyti sistemos mokymosi procesus, paremtus tiriamaisiais duomenimis. Šiame etape galima taikyti klasifikavimo metodus paruošiant modelį ir atliekant naujai gautų duomenų prognozes. Etapą galima išplėsti papildomais vizualizavimo sprendimais, padedančiais paaiškinti ir suprasti duomenis, o ne tik pateikti pačius duomenis, t. y. statistinius duomenis, sistemos mokymosi rezultatus ir pan.



3 pav. Atsiliepimams apie prekes apdoroti skirtas darbų sekos modelis

Duomenų apdorojimas visiems pateiktiems algoritmams yra toks pat, todėl šiame poskyryje aprašytas unifikuotas duomenų išankstinio apdorojimo procesas. Duomenų išankstinio apdorojimo procesas yra pirmasis ir pradinis etapas, į kurį turi būti atsižvelgta duomenų analizės sprendimų kūrimo pradžioje. Visi modelio žingsniai plačiau paaiškinami tolesnėse šio skyriaus dalyse.

Duomenų išgavimas

Pagrindinis šio etapo tikslas yra atrinkti tik reikiamus ir susijusius duomenų laukus, siekiant tinkamai apdoroti duomenis ir optimizuoti atminties panaudojimą. Šis etapas atliktas tokiais žingsniais:

- Įkeliamas duomenų rinkinio failas, kuriame yra atsiliepimai apie prekes.
- Atrinkami duomenys pagal antraštę (*overall*), autorių (*author*), autorizacijos numerį (*id*), atsiliepimo tekstą (*reviewText*) ir pan.

- Iš įvesties duomenų rinkinių atrinkami tik reikiami *Overall* ir *reviewText* laukai.
- Atrenkamas vienodas skaičius atsiliepimų apie prekes kiekvienai klasei, t. y. naudojant asimetrišką (angl. *skewness*) duomenų išsidėstymo metodą.

Atsiliepimų apie prekes tekstų parengimas

Pagrindinis šio etapo tikslas yra parengti *reviewText* laukus duomenų požymiams gauti. Šiame etape pritaikytas duomenų požymių parinkimo metodai:

- Pašalinami skyrybos ženklai ir kiekvienas žodis segmentuojamas atsižvelgiant į tarp žodžių esančius tarpus.
- Pašalinami nereikšmingi žodžiai, tokie kaip anglų kalbos artikeliai *a* ir *the* (šie artikeliai dažnai naudojami tekste, tačiau juose nėra jokios informacijos, kuri būtų reikšminga ruošiant duomenų modelį). Nereikšmingų žodžių pašalinimas nėra būtinas, kadangi atsiliepimai apie prekes yra trumpi ir naudojant tokias metodikas, kaip atvirkštinis žodžių dažnis (TF-IDF), jų įtaka gali būti sumažinta.
- Visos didžiosios raidės pakeičiamos į mažąsias (angl. *lowercasing*).
- Žodžių grąžinamas į jų bendrinę formą. Šiam tikslui taikomas Porter stemmer metodas. Bendrinės žodžio formos grąžinimas nebūtinai pagerina klasifikavimo kokybę, todėl ši įtaka turėtų būti eksperimentiškai ištirta.

Žodžių krepšeliai

N-gramų metodas yra taikomas konstruojant žodžių krepšelius pagal *n* ilgio iš eilės einančių žodžių seką:

- Padalijami sakiniai į žodžius ir žodžiai sugrupuojami pagal iš anksto apibrėžtus *n*-gramų junginius. Žodžių krepšeliai (unigramos, bigramos ir trigramos ar jų deriniai) sudaromi iš atsiliepimų apie prekes tekstų, kurie jau buvo apdoroti ankstesniuose etapuose, pagal pasirinktą *n*-gramų modelį. *N*-gramos yra sudaromos iš *n* vienetų (simbolių ar žodžių) tam tikros teksto sekos.
- Naudojami tęstiniai tekstai vietoj *n*-gramų derinių kūrimo segmentuojant sakinius. Taip yra todėl, kad klasifikatoriaus užduotis nėra suprasti sakinio reikšmę, tačiau sakinys sukuria klasifikatoriui įvesties duomenis su visais jų požymiais. Todėl klasifikatorius sukuria modelį, priskiriantį duomenis tam tikrai klasei taip tiksliai, kaip įmanoma.
- Taikant kalbos dalies žymėjimo metodą (angl. *Part-of-the-speech*), kiekvienam žodžiui pridedama žyma, jei tai būdvardis, daiktavardis, veiksmažodis, ar kt.

Teksto klasifikavimas

Šis etapas vykdytas tokiais žingsniais:

- Duomenų rinkinys dalijamas į dvi grupes, apytikriai 90 % duomenų paliekami apmokymui, ir 10 % duomenų – testavimui.
- *N* kartų kryžminės validacijos metodo naudojimas. Šiuo metodu duomenys suskirstomi į *n* skirtingų, bet vienodo dydžio pogrupių. Po to kiekvienas pogrupis naudojamas testavimo etape. Likę pogrupiai kombinuojami mokymo etape klasifikatoriaus mokymuisi. Tekstinių duomenų apmokymas ir testavimas

atliekamas pagal pasirinktą klasifikavimo metodą, naudojant 10-ties kartų kryžminę validaciją.

- Naudoti klasifikatoriai: paprastasis Bajesas, atsitiktinis miškas, sprendimų medis, atraminių vektorių mašinų metodas su tiesinio branduolio ir stochastinio gradientinio nusileidimo optimizavimo algoritmu, logistinė regresija su ribotos atminties Broyden–Fletcher–Goldfarb–Shanno optimizavimo algoritmu ir daugiasluoksnio perceptrono algoritmas.
- Klasifikavimo metodai naudoti su jų standartiniais hiperparametrais, sukonfigūruotais Apache Spark v1.6.2 MLlib bibliotekoje, išskyrus nurodant požymių vektoriaus ilgį (3000), medžio dydį (50) ir gylį (30) – pastarieji pritaikyti pagal duomenų dydį ir apribojimus, susijusius su atsitiktinio miško metodo ir kompiuterinių resursų naudojimu. Tam tikri parametrai, kaip įvestis (200), paslėptasis sluoksnis 1 (20), paslėptasis sluoksnis 2 (10) ir išvestis (5), buvo naudojami daugiasluoksnio perceptrono klasifikatoriaus atveju.

3. EKSPERIMENTAVIMAS IR REZULTATAI

Šiame skyriuje pristatomi lyginamojo tyrimo, įskaitant eksperimento planavimą, rezultatai. Palyginimas yra paremtas duomenų požymių parinkimo deriniais (*n*-gramų, kalbos dalių žymėjimo, (atvirkštiniais) žodžių dažniais, žodžių įterpimą) ir klasifikatoriais (Bajeso, atsitiktinio miško, sprendimų medžio, atraminių vektorių mašinų, logistinės regresijos, daugiasluoksnio perceptrono) skirtiems daugiaklasiams tekstiniams duomenims klasifikuoti, siekiant pagerinti klasifikavimo tikslumą, naudojant dvi didelės apimties atsiliepimų apie prekes duomenų aibes.

Eksperimento planavimas

Eksperimento planavimas yra paremtas atrinktų Bajeso, atsitiktinio miško, sprendimų medžio, atraminių vektorių mašinų, logistinės regresijos ir daugiasluoksnio perceptrono metodų vertinimu daugiaklasiams tekstiniams duomenims klasifikuoti, įskaitant ir duomenų požymių atranką: *n*-gramas, kalbos dalis, (atvirkštinį) žodžių dažnį, žodžių įterpimą.

Hipotezių formulavimas ir kintamųjų pasirinkimas

Hipotezės ir probleminiai klausimai paprastai padeda tinkamai apibrėžti ir suplanuoti eksperimentinę analizę. Norint atsakyti į disertacijos pradžioje suformuluotus klausimus ir įgyvendinti užsibrėžtus tikslus, reikia atlikti eksperimentą.

- Formuluojama hipotezė: duomenų požymių parinkimo junginiai (n-gramų junginiai) gali padidinti klasifikavimo tikslumą.
- Remdamasis šia hipoteze, šios disertacijos autorius prognozuoja, kad duomenų požymių parinkimo derinio naudojimas leis pasiekti didesnį klasifikavimo tikslumą taikant pasirinktus atsiliepimų apie prekes duomenų klasifikavimo metodus, lyginant su rezultatais, gaunamais taikant bazinius klasifikavimo metodus.
- Atliekant mokslinį eksperimentą, svarbu pasirinkti keletą nepriklausomų kintamųjų kaip veiksnį, kuris bus keičiamas eksperimento metu. Priklausomi kintamieji kaip veiksnys prognozuojamai keisis. Apibrėžkime funkciją f ir išvestį A_i kaip formalią klasifikavimo tikslumo išraišką:

$$A_i = f\ (C_{metodai}, Požymiai, NLP_{technikos}, D_i,) \quad (1)$$

Nepriklausomi kintamieji šiam eksperimentui apibrėžiami taip:

- Klasifikavimo metodai *($C_{metodai}$)*: paprastasis Bajesas, atsitiktinis miškas, sprendimų medis, atraminių vektorių mašinų metodas, logistinė regresija ir daugiasluoksnis perceptronas.
- Duomenų požymių atranka *(Požymiai)*: *n*-gramos, kalbos dalys, žodžių dažnis, atvirkštinis žodžių dažnis, žodžių įterpimas.
- Triukšmo mažinimas ($NLP_{technikos}$): žodžių skaidymas (segmentavimas), nereikšmingi žodžiai, raidžių keitimas mažosiomis raidėmis, žodžių normalizavimas (į pagrindinę žodžio formą) ir kt.

- Duomenų rinkiniai *(D_i)*: klasifikavimo eksperimentai vykdyti su dviem nepriklausomais duomenų rinkiniais (A ir B), kuriuos naudojant statistiškai išmatuotas klasifikavimo tikslumas. Taip pat buvo išmatuoti: klaidų lygis (angl. *error rate*), preciziškumas (angl. *precision*), prisiminimas (angl. *recall*) ir F1 rodiklis (angl. *F1-measurement*).

Eksperimentą sudaro 4 eksperimentiniai ciklai, kurių kiekviename naudoti nepriklausomi kintamieji (pvz., klasifikavimo metodai, duomenų požymių parinkimas ir kt.) naudoti taip, kaip aprašyta 1 lentelėje. 2-ame, 3-ame ir 4-ame eksperimentiniame cikle taikyti tik 3–4 atrinkti geriausi klasifikavimo metodai.

1 lentelė. Eksperimentiniai ciklai

Kintamieji / Ciklas	$C_{metodai}$	$Požymiai$	$NLP_{technikos}$
1-as	Bajesas, atsitiktinis miškas, sprendimų medis, atraminių vektorių mašinos, logistinė regresija, daugiasluoksnis perceptronas	1. Žodžių dažniai (išskyrus daugiasluoksnį perceptroną naudojant word2vect) 2. Unigrama, bigrama ir trigrama 3. *N*-gramų junginiai	Segmentavimas, mažosios raidės, nereikšmingų žodžių šalinimas, žodžio grąžinimas į pagrindinę formą
2-as	Bajesas, atraminių vektorių mašinos, logistinė regresija	1. Atvirkštinis žodžių dažnis 2. Unigrama ir *n*-gramų junginiai	Segmentavimas, mažosios raidės, nereikšmingų žodžių šalinimas, žodžio grąžinimas į pagrindinę formą
3-as	Bajesas, atraminių vektorių mašinos, logistinė regresija, daugiasluoksnis perceptronas	1. Atvirkštiniai žodžių dažniai (išskyrus daugiasluoksnį perceptroną su word2vect) 2. Unigrama ir *n*-gramų junginiai 3. Kalbos dalių žymėjimas	Segmentavimas, mažosios raidės, nereikšmingų žodžių šalinimas, žodžio grąžinimas į pagrindinę formą
4-as	Paprastasis Bajesas, atraminių vektorių mašinos, logistinė regresija	1. Atvirkštinis žodžių dažnis 2. Unigrama ir *n*-gramų junginiai	Segmentavimas, mažosios raidės, netaikant nereikšmingų žodžių šalinimo, žodžio grąžinimas į pagrindinę formą

Eksperimento kontrolė ir duomenų rinkimas

Atliekant mokslinį eksperimentą, svarbu apibrėžti lyginimo standartą, kad visi eksperimentai būtų traktuojami vienodai:

- Šiame eksperimente klasifikavimo tikslumas naudojamas kaip pagrindinė kontrolės priemonė ir priklausomas kintamasis. Taip pat eksperimentuose matuoti klaidų lygis, preciziškumas, prisiminimas ir F1 rodiklis.
- Lyginamasis daugialypių klasifikavimo metodų – Bajeso, atsitiktinio miško, sprendimų medžio, atraminių vektorių mašinų su tiesiniu branduoliu ir stochastinio gradientinio nusileidimo (angl. Stochastic Gradient Descent) optimizavimo

algoritmu, logistinės regresijos su ribotos atminties Broyden–Fletcher–Goldfarb–Shanno optimizavimo algoritmu bei daugiasluoksnio perceptrono – klasifikavimo tikslumas siejamas su atsiliepimų apie prekes skaičiumi bei *n*-gramų junginiais atitinkamai abiem duomenų rinkiniams.
- Duomenų rinkiniai suskirstyti į dvi grupes, apytikriai priskiriami 90 % duomenų apmokymui ir 10 % duomenų testavimui.
- Klasifikavimo metodai naudoti su numatytaisiais parametrais, esančiais Apache Spark v1.6.2 MLlib bibliotekoje.

Visi eksperimentai atlikti tuo pačiu būdu, siekiant užtikrinti, kad surinkti duomenys atitiktų tas pačias sąlygas visose eksperimentinėse grupėse. Tų pačių sąlygų užtikrinimas visiems eksperimentams reiškia tai, kad:
- Naudoti vienodo dydžio paskirstyti atsiliepimų apie prekes duomenų rinkiniai kiekvienoje klasėje.
- Naudotas 10-ties kartų kryžminės validacijos metodas, siekiant validuoti klasifikavimo eksperimentų rezultatus ir išvadas.
- Duomenys apie klasifikavimo tikslumą ir kitus iš anksto apibrėžtus matavimus susisteminti diagramose ir lentelėse.
- Naudota ta pati programinė ir infrastruktūrinė įranga pristatyta 2 skyriuje.

Rezultatų apžvalga

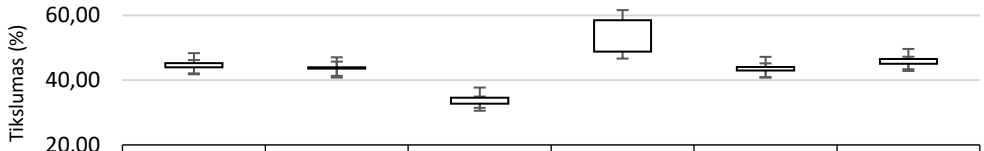
Eksperimentų rezultatai parodė, kad net jei klasifikavimo metodai yra plačiai žinomi ir taikomi tekstiniams bei kito tipo duomenims klasifikuoti naudojant sistemos mokymo įrankius, jų klasifikavimo tikslumas koreliuoja su įvesties duomenų pakitimais naudojamais klasifikavimo modelyje, pavyzdžiui, taikant duomenų požymių parinkimo ir natūralios kalbos apdorojimo technikas, tokias kaip triukšmo mažinimas. Tai reiškia, klasifikavimo tikslumas labai priklauso nuo pasirinktų požymių ir *n*-gramų junginių.

Tyrimo rezultatai rodo, kad daugeliu atvejų pasirinktos *n*-gramų junginiai leidžia 1–5 % padidinti klasifikavimo tikslumą, tačiau šis pokytis nėra reikšmingas. Be to, tokie metodai, kaip asimetriškas duomenų išdėstymas (angl. *skewness*) ir kryžminė validacija yra klasikiniai ir universalūs visiems eksperimento ciklams. Kaip minėta anksčiau, asimetriško duomenų išdėstymo metodas padeda surinkti po vienodą skaičių atsiliepimų apie prekes kiekvienoje klasėje, o kryžminė validacija užtikrina klasifikavimo eksperimentų rezultatų ir išvadų patikimumą. Visiems metodams ir eksperimentiniuose cikluose naudoti šie triukšmo mažinimo įrankiai: žodžių segmentavimas, nereikšmingų žodžių šalinimas (išskyrus 4-ą ciklą), skyrybos ženklų šalinimas, visų didžiųjų raidžių pakeitimas mažosiomis, žodžių grąžinimas į pagrindinę formą.

Nė viename eksperimento cikle nebuvo vertinama nei atsiliepimų apie prekes kokybė, nei jų rašybos tikslumas naudojant rašybos korekcijos įrankius, nevertinta ir šypsenėlių įtaka (angl. *emoticons*). Eksperimentiškai nustatyta, kad žodžių rašybos tikslumo tikrinimas kompiuteriniu rašybos tikrinimo įrankiu dėl didelio žodžių skaičiaus įprastai užima daug daugiau laiko, nei tikėtasi, tad rašybos tikrinimo buvo nuspręsta atsisakyti. Testavimo metu rezultatai parodė, kad rašybos tikrinimas gali padidinti klasifikavimo tikslumą 1–2 %, kai taikomas Bajeso klasifikavimo metodas. Tačiau stilistines ar leksikos klaidas yra sunkiau taisyti, nes jos yra rašytinės šnekamosios kalbos formos rezultatas. Be to, ne kiekvienu atveju šie taisymai įmanomi, kadangi joms taisyti

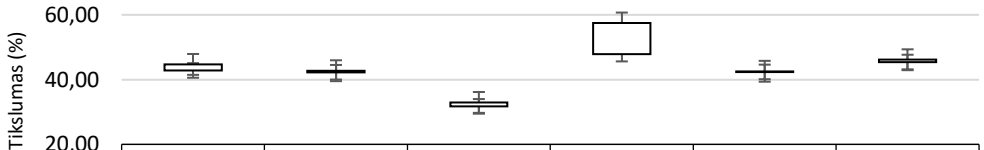
nėra tikslių kalbos korekcijos įrankių. Vis dėlto pastaruoju metu vis plačiau naudojamas dirbtinis neuroninis tinklas, leidžiantis išvengti bet kokio tekstinių duomenų keitimo, kadangi toks klasifikatorius geba geriau įsiminti tokius duomenis, kokie jie iš tiesų yra. Idealiu atveju, kai naudojamas dirbtinis neuroninis tinklas, tekstiniai duomenys neturėtų būti papildomai apdorojami natūralios kalbos apdorojimo įrankiais.

4 pav. pateikti geriausiai veikiančių *n*-gramų (palyginus unigramą su unigrama / bigrama ir unigrama / bigrama / trigrama) vidutinio klasifikavimo tikslumo rezultatai gauti 1-ame eksperimentiniame cikle. Rezultatai rodo, kad vidutinės klasifikavimo tikslumo reikšmės klasifikuojant duomenis su Bajeso, atsitiktinio miško ir atraminių vektorių mašinų metodais yra panašios (duomenų rinkinys A: min unigrama: 42,99–44,01 %, max unigrama, bigrama ir trigrama: 43,93–45,22 %; duomenų rinkinys B: min unigrama: 42,27–42,82 %, max unigrama, bigrama ir trigrama: 42,57–44,69 %). Klasifikuojant duomenis Bajeso metodu, pasiektas 1–2 % didesnis klasifikavimo tikslumas, lyginant su rezultatais, gautais pritaikius atsitiktinio miško ir atraminių vektorių mašinų metodus, tačiau šis skirtumas nėra statistiškai reikšmingas.



A)

	Bajesas	Atsitiktinis miškas	Sprendimų medis	Logistinė regresija	Atraminių vektorių mašinos	Daugiasl. perceptronas
unigrama	44,01	43,53	32,74	48,86	42,99	45,04
uni-/ bi-/ trigrama	45,22	43,93	34,58	58,5	44,06	46,51



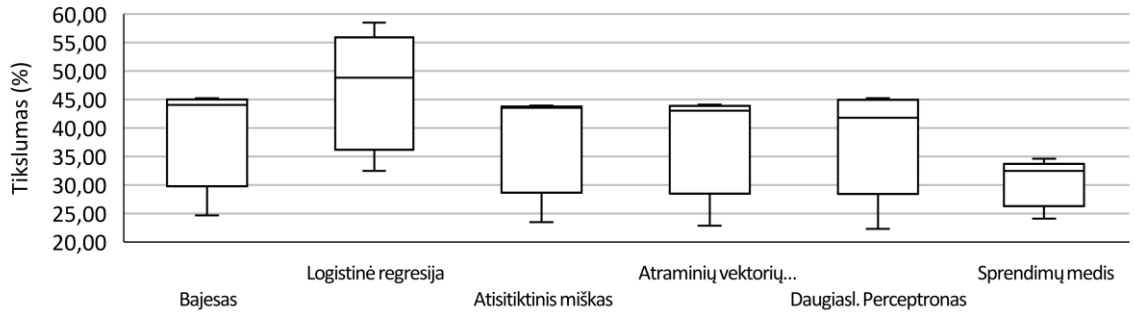
B)

	Bajesas	Atsitiktinis miškas	Sprendimų medis	Logistinė regresija	Atraminių vektorių mašinos	Daugiasl. perceptronas
unigrama	42,82	42,27	31,73	47,87	42,37	45,42
uni-/ bi-/ trigrama	44,69	42,74	32,95	57,48	42,57	46,13

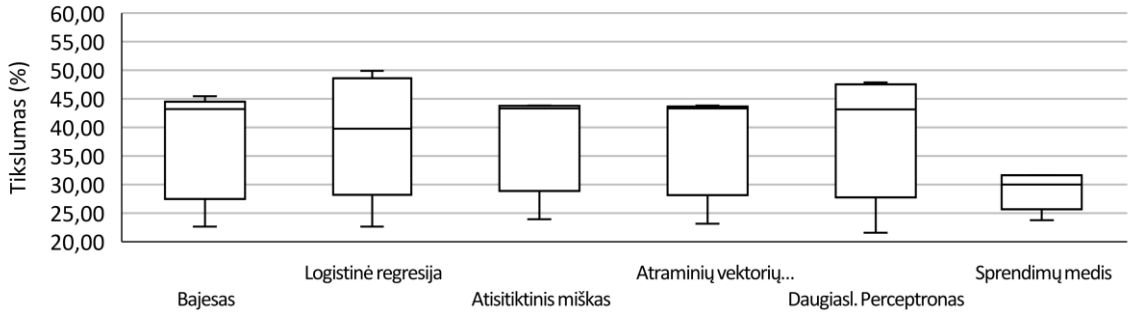
4 pav. Vidutinis klasifikavimo tikslumas 1-ame eksperimentiniame cikle

1-o eksperimentinio ciklo rezultatai rodo, kad naudojant *n*-gramų ir žodžių dažnių parinkimo požymius, logistinės regresijos daugiaklasis klasifikavimo metodas pasiekė didžiausią vidutinį atsiliepimų apie prekes klasifikavimo tikslumą (min 32,43 %, max 58,50 %), lyginant su kitų metodų – Bajeso, atsitiktinio miško, sprendimų medžio ir atraminių vektorių mašinų – klasifikavimo tikslumu. Naudojant bazinį *(angl. baseline)* (~20 %) ir sprendimų medžio klasifikatorius, gautos mažiausios vidutinės klasifikavimo tikslumo reikšmės (min trigrama: 24,10 %, max unigrama, bigrama ir trigrama: 34,58 %). Didžiausias vidutinis klasifikavimo tikslumas pasiektas logistinės regresijos metodu, pranokusiu kitus metodus, kai taikomi žodžių dažnio ir *n*-gramų junginiai (min: 57 %, max: 58 %). O logistinės regresijos su unigrama klasifikavimo tikslumo rezultatai yra prastesni (min: 47 %, max: 48 %). Ekvivalentūs rezultatai gauti su abiem duomenų rinkiniais.

1-o eksperimentinio ciklo rezultatai rodo, kad didinant mokymo duomenų rinkinio dydį nuo 5000 iki 75000 atsiliepimų apie prekes vienoje klasėje, klasifikavimo tikslumas nereikšmingai (1–2 %) didėja, kai naudojami Bajeso, atsitiktinio miško, atraminių vektorių mašinų ir logistinės regresijos klasifikavimo metodai. Šie rezultatai rodo, kad mokymo duomenų rinkinio dydis, lygus 5000 atsiliepimų apie prekes vienoje klasėje, yra pakankamas. Tik logistinės regresijos klasifikatoriaus naudojimas 2-ame ir 3-ame eksperimentiniuose cikluose, kai mokymo duomenų rinkinio dydis siekia nuo 5000 iki 45000 atsiliepimų apie prekes vienoje klasėje, nulėmė klasifikavimo tikslumo didėjimą (8–9 %), lyginant su duomenų rinkinio, kuriame yra 5000 atsiliepimų, tikslumu. Tai reiškia, kad duomenų rinkinio didinimas padeda pasiekti didesnį klasifikavimo tikslumą. Kai taikomi kiti metodai, klasifikavimo tikslumo padidėjimas nėra reikšmingas, ir pats klasifikavimo tikslumas daugiau sietinas su duomenų požymių parinkimo būdais, tokiais kaip *n*-gramos, (atvirkštinis) žodžių dažnis ir kalbos dalių žymėjimas.



A)



B)

5 pav. Vidutinės klasifikavimo tikslumo absoliučiosios reikšmės 1-ame eksperimentiniame cikle

5 pav. pateiktos absoliučiosios reikšmės rodo, kad didžiausias klasifikavimo tikslumas pasiekiamas naudojant logistinės regresijos klasifikatorių su siūlomomis *n*-gramų (unigrama, bigrama ir trigrama) junginiais, lyginant su unigramų ir kitų klasifikavimo metodų, eksperimentiškai analizuotų 1-ame cikle, klasifikavimo tikslumu. Tai reiškia, tiesinės logistinės regresijos metodui būdinga tai, kad klasifikavimo tikslumo reikšmės yra mažiau stabilios ir labiau pasiskirsčiusios. Išskyrus logistinę regresiją, kitų 1-ame eksperimentiniame cikle analizuotų metodų veikimas yra patikimesnis, o vidutinės klasifikavimo tikslumo reikšmės yra mažiau pasiskirsčiusios (žr. 4 pav.).

A)

	Bajesas	Logistinė regresija	Atraminių vektorių mašinos
unigrama	43,67	45,47	42,99
uni-/bigrama	44,25	45,03	46,2
uni-/ bi-/ trigrama	44,30	44,15	46,25

B)

	Bajesas	Logistinė regresija	Atraminių vektorių mašinos
unigrama	43,67	45,46	45,74
uni-/bigrama	44,3	45,07	46,25
uni-/ bi-/ trigrama	44,32	44,13	46,24

6 pav. Vidutinis klasifikavimo tikslumas 2-ame eksperimentiniame cikle

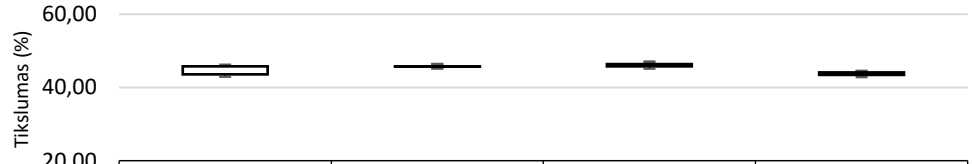
6 pav. pateikiamos vidutinės klasifikavimo tikslumo reikšmės 2-ame eksperimentiniame cikle. Bajeso (duomenų rinkinys A: min unigrama: 43,67 %, max unigrama, bigrama ir trigrama: 44,30 %; duomenų rinkinys B: min unigrama: 43,67 %, max unigrama, bigrama ir trigrama: 44,32 %) ir atraminių vektorių mašinų (duomenų rinkinys A: min unigrama: 42,99 %, max unigrama, bigrama ir trigrama: 46,25 %; duomenų rinkinys B: min unigrama: 45,74 %, max unigrama, bigrama ir trigrama: 46,24 %) klasifikatoriai su siūlomomis *n*-gramų (unigrama, bigrama ir trigrama) junginiais pasiekė didžiausią vidutinį klasifikavimo tikslumą, lyginant su tikslumu, gautu naudojant unigramas. O logistinės regresijos metodas pasiekė didžiausią klasifikavimo tikslumą naudojant unigramas (duomenų rinkinys A: max unigrama: 45,47 %, A: min unigrama, bigrama ir trigrama: 44,15 %; duomenų rinkinys B: max unigrama: 45,46 %, min unigrama, bigrama ir trigrama: 44,13 %), o naudojant *n*-gramų (unigrama, bigrama ir trigrama) junginius, klasifikavimo tikslumas yra mažesnis. Didžiausią klasifikavimo tikslumą 2-ame eksperimentiniame cikle pasiekė atraminių vektorių mašinų metodas (duomenų rinkinys A: max unigrama, bigrama ir trigrama 46,25 %; duomenų rinkinys B: max unigrama, bigrama ir trigrama 46,24 %) su *n*-gramų (unigrama, bigrama ir trigrama) junginiais.

2-o eksperimentinio ciklo, kuriame naudoti *n*-gramų junginiai ir atvirkštinis žodžių dažnio duomenų parinkimo požymiai, rodo, kad atraminių vektorių mašinų daugiaklasis klasifikavimo metodas pasiekė didžiausią klasifikavimo tikslumą (min: 45 %, max: 46 %) klasifikuojant atsiliepimų apie prekes duomenis, lyginant su Bajeso ir logistinės regresijos klasifikavimo metodų tikslumu. Naudojant logistinės regresijos metodą, kai duomenų rinkinio dydis lygus 5000, klasifikavimo tikslumas sumažėjo (min: 38 %, max: 41 %), tačiau duomenų rinkiniui padidėjus iki 10000 atsiliepimų, klasifikavimo tikslumas padidėjo (46–47 %). Su abiem duomenų rinkiniams gauti ekvivalentūs rezultatai.

A)

	Bajesas	Logistinė regresija	Atraminių vektorių mašinos	Daugiasl. perceptronas
unigrama	43,57	45,73	45,86	43,12
uni-/bigrama	44,71	45,68	46,51	43,35
uni-/ bi-/ trigrama	45,46	45,25	46,39	43,58

B)



	Bajesas	Logistinė regresija	Atraminių vektorių mašinos	Daugiasl. perceptronas
unigrama	43,54	45,77	45,74	43,43
uni-/bigrama	44,6	45,85	46,52	43,85
uni-/ bi-/ trigrama	45,77	45,74	46,37	44,11

7 pav. Vidutinis klasifikavimo tikslumas 3-iame eksperimentiniame cikle

3-čio eksperimentinio ciklo (7 pav.), kuriame naudotos *n*-gramų, atvirkštinio žodžių dažnio ir kalbos dalių žymėjimo požymiai, rezultatai rodo, kad *atraminių vektorių mašinų* daugiaklasis klasifikavimo metodas, naudotas klasifikuoti atsiliepimų apie prekes duomenims, pasiekė didžiausią klasifikavimo tikslumą (min: 45 %, max: 47 %) lyginant su Bajeso ir logistinės regresijos klasifikavimo metodų tikslumu. Su abiem duomenų rinkiniais gauti ekvivalentūs rezultatai.

A)

	Bajesas	Logistinė regresija	Atraminių vektorių mašinos
unigrama	44,79	47,89	48,02
uni-/bigrama	47,26	48,4	49,06
uni-/ bi-/ trigrama	47,44	46,37	48,92

B)

	Bajesas	Logistinė regresija	Atraminių vektorių mašinos
unigrama	42,13	46,81	46,58
uni-/bigrama	43,78	46,86	45,89
uni-/ bi-/ trigrama	44,18	46,27	45,58

8 pav. Vidutinis klasifikavimo tikslumas 4-ame eksperimentiniame cikle

Kai kurie autoriai teigia, kad nereikšmingų žodžių šalinimas ir žodžių grąžinimas į pagrindinę formą nebūtinai pagerina klasifikavimo kokybę, todėl buvo tikslinga šiuos

teginius patikrinti eksperimentiškai. Tai ir buvo atlikta 4-ame eksperimentiniame cikle (8 pav.). Rezultatai parodė, kad atvirkštinio žodžių dažnio ir nereikšmingų žodžių šalinimo naudojimas taikant *Bajeso* (duomenų rinkinys A: min unigrama: 44,79 %, max unigrama, bigrama ir trigrama: 47,44 %; duomenų rinkinys B: min unigrama: 42,13 %, max unigrama, bigrama ir trigrama: 44,18 %) ir atraminių vektorių mašinų (duomenų rinkinys A: min unigrama: 48,02 %, max unigrama ir bigrama: 49,06 %; duomenų rinkinys B: min unigrama, bigrama ir trigrama: 45,58 %, max unigrama: 46,58 %) metodus kartu su unigrama ir *n*-gramų junginiais pasiekė apytikriai 2 % didesnį klasifikavimo tikslumą, lyginant su klasifikavimo tikslumu, gautu 2-ame eksperimentiniame cikle (kuomet buvo pašalinti nereikšmingi žodžiai). Tuo tarpu gauti rezultatai naudojant logistinės regresijos metodą išliko panašūs lyginant su rezultatais gautais 2-ame eksperimentiniame cikle (duomenų rinkinys A: min unigrama, bigrama ir trigrama: 46,37 %, max unigrama ir bigrama: 48,4 %; duomenų rinkinys B: min unigrama, bigrama ir trigrama: 46,27 %, max unigrama ir bigrama: 46,86 %).

Eksperimentų rezultatai, kuomet nebuvo naudotas žodžių grąžinimas į pagrindinę žodžio formą, gauti ne iš anksto apibrėžtų eksperimentinių ciklų metu. Todėl šie eksperimentai vykdyti tik su vienu duomenų rinkiniu DS4 (30000 vienoje klasėje). Vidutinis klasifikavimo tikslumas yra apytikriai 2 % mažesnis naudojant Bajeso, logistinės regresijos ir atraminių vektorių mašinų daugiaklasius klasifikavimo metodus kartu su unigramomis ir *n*-gramų junginius, nei klasifikavimo tikslumas, gautas 2-ame eksperimentiniame cikle (kuomet buvo naudotas žodžių grąžinimas į pagrindinę žodžio formą).

2 lentelė. Geriausiai veikiantys metodai su parinktais duomenų požymiais atsiliepimams apie prekes klasifikuoti

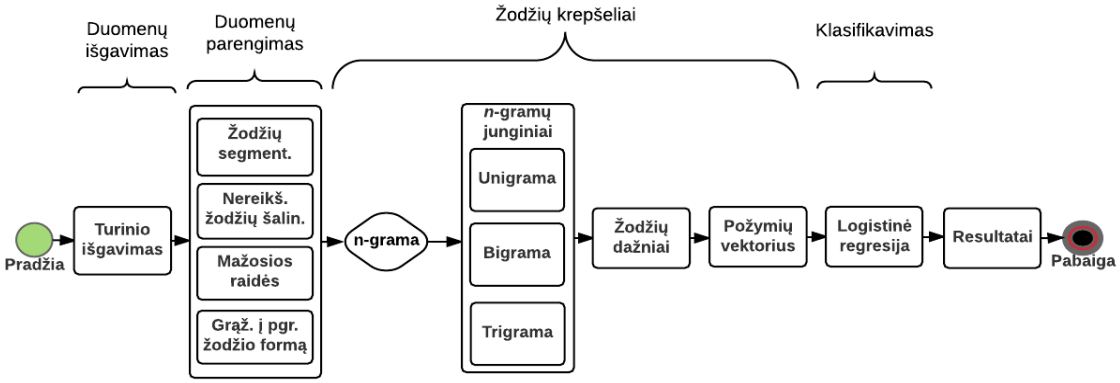
Ciklas	Didžiausias vidutinis tikslumas (duomenų rinkinys A)	Didžiausias vidutinis tikslumas (duomenų rinkinys B)	Klasifikatorius
1	**58,48 %**	**57,62 %**	**Logistinė regresija**
	45,18 %	45,80 %	Daugiasluoksnis perceptronas
	44,06 %	44,00 %	Atraminių vektorių mašina
2	46,25 %	46,26 %	Atraminių vektorių mašinos
	45,46 %	45,46 %	Logistinė regresija
3	45,73 %	46,53 %	Atraminių vektorių mašinos
	46,51 %	45,58 %	Logistinė regresija
4	49,06 %	46,58 %	Atraminių vektorių mašinos
	48,4%	46,86%	Logistinė regresija

Apibendrinant gautus rezultatus galima teigti, kad geriausiai veikiantis trumpųjų tekstinių žinučių (atsiliepimų apie prekes) daugiaklasis klasifikavimo metodas yra logistinė regresija, taikoma su tokiais parinktais duomenų požymiais, kaip žodžių dažnis ir *n*-gramų junginiais (unigrama, bigrama ir trigrama). Geriausiai veikiančių klasifikavimo metodų su parinktais duomenų požymiais rezultatai, gauti kiekviename eksperimentiniame cikle, pateikiami 2 lentelėje.

Siūlomas klasifikavimo metodas

Pagrindinis šios disertacijos siekis buvo išanalizuoti klasikinius klasifikavimo metodus ir sukurti duomenų požymių parinkimo metodą naudojant *n*-gramų junginius ir žodžių dažnius, šį metodą taikant daugiaklasiams trumpiems tekstiniams duomenims klasifikuoti. Pasiūlyto metodo schema su geriausiai veikiančiais duomenų požymių parinkimo metodais yra pavaizduota 9 pav. Schemoje esantys etapai, siūlomi kaip šio metodo dalis kartu su duomenų požymių parinkimu, apimančiu *n*-gramų junginius ir žodžių dažnius. Išgavus duomenis, duomenų parengimo etape turi būti pašalinti skyrybos ženklai, segmentuojant žodžiai atskiriami vienas nuo kito pagal esančius tarpus tarp žodžių, pašalinami nereikšmingi žodžiai. Artikeliai „*a*“ ir „*the*“ dažnai naudojami bet kokiame tekste (anglų k.), tačiau ruošiant duomenų modelį juose nėra jokios reikšmingos informacijos. 4-ame eksperimentiniame cikle nustatyta, kad nereikšmingų žodžių šalinimas nėra būtinas. Šių žodžių įtaka gali būti sumažinta naudojant tokias technikas, kaip atvirkštinis žodžių dažnis. Didžiosios raidės pakeičiamos mažosiomis iš karto po nereikšmingų žodžių pašalinimo. Likę žodžiai grąžinami į pagrindinę žodžio formą (angl. *stemming*) sumažina išvestinių žodžių kiekį tekste. Žodžių grąžinimas į pagrindinę žodžio formą nebūtinai pagerina klasifikatoriaus kokybę, tačiau eksperimentinė analizė parodė, kad įtaką klasifikavimo tikslumui yra reikšminga.

Atlikta Bajeso, atsitiktinio miško, sprendimų medžio, atraminių vektorių mašinų, logistinės regresijos ir daugiasluoksnio perceptrono metodų, taikomų daugiaklasiams tekstiniams duomenims klasifikuoti, lyginamoji analizė. Tyrimo rezultatai rodo, kad logistinės regresijos daugiaklasių tekstinių duomenų klasifikavimo metodas, taikytas klasifikuojant atsiliepimus apie prekes duomenis 1-ame eksperimentiniame cikle, pasiekė didžiausią (mažiausiai 32–33 %, daugiausiai 57–58 %) klasifikavimo tikslumą, lyginant su Bajeso, atsitiktinio miško, sprendimų medžio ir atraminių vektorių mašinų metodais.



9 pav. Siūlomo metodo schema

Taikant sprendimų medžio metodą, gautos mažiausios vidutinės klasifikavimo tikslumo reikšmės (mažiausiai su trigrama: 24,10 %, daugiausiai su unigrama, bigrama ir trigrama: 34,58 %). Eksperimentų rezultatai parodė, kad Bajeso klasifikavimo metodas, taikytas atsiliepimų apie prekes duomenims 1-ame eksperimentiniame cikle, pasiekė 1–2 % didesnį vidutinį klasifikavimo tikslumą, nei atsitiktinio miško ir atraminių vektorių mašinų metodai, tačiau šis skirtumas nėra reikšmingas statistiškai. Remiantis

lyginamosios analizės rezultatais, galima teigti, jog bendrasis klasifikavimo tikslumas jungiant vieną, dvi ir tris gramas didina vidutinį klasifikavimo tikslumą (vidutiniškai ~10 %), tačiau šios reikšmės nėra reikšmingos lyginant visų klasifikavimo metodų rezultatus gautus naudojant unigramą. Tyrimas atskleidė, kad didinant apmokymo duomenų rinkinio dydį nuo 5000 iki 75000 atsiliepimų apie prekes vienoje klasėje sąlygoja nereikšmingą (1–2 %) klasifikavimo tikslumo padidėjimą taikant Bajeso, atsitiktinio miško ir atraminių vektorių mašinų metodus. Šie rezultatai rodo, kad apmokymo duomenų rinkinio dydis lygus 5000 atsiliepimų apie prekes vienoje klasėje, yra tinkamas visiems analizuotiems klasifikavimo metodams, išskyrus logistinės regresijos metodą, taikytą su atvirkštinio žodžių dažnio duomenų parinkimo požymiu (2-ame eksperimentiniame cikle). Visiems klasifikavimo metodams tikslumas koreliuoja nuo *n*-gramų junginių parinkimo ir taikymo, tačiau logistinės regresijos ir daugiasluoksnio perceptrono metodų atveju ir nuo duomenų rinkinio dydžio.

Tyrimo rezultatai atskleidė, kad logistinė regresija pasiekia geresnių rezultatų nei atraminių vektorių mašinų ir daugiasluoksnio perceptrono metodai, kai taikoma žodžių dažnių parinkimo funkcija. Vis dėlto abu pastarieji metodai gali būti rekomenduojami trumpiems tekstams klasifikuoti, nors siekiant didesnio klasifikavimo tikslumo, būtų tikslinga optimizuoti pasirinktus parametrus arba naudoti kitus netiesinius atraminių vektorių mašinų metodo branduolius.

4. BENDROSIOS IŠVADOS

1. Šioje disertacijoje gauti rezultatai ir atlikti lyginimai rodo, kad debesų kompiuterijos technologijos integruojančios didelės apimties duomenų analitikos karkasus, gali būti sėkmingai naudojamos atminčiai imliems klasifikatorių mokymo algoritmams.
2. Atlikus klasifikavimo algoritmų lyginimą, galima teigti, kad klasifikavimo tikslumas, paremtas *n*-gramų (unigramos, bigramos ir trigramos) junginiais ir jų dažniu paremtu duomenų požymių sudarymu, didina vidutinį trumpų tekstų, tokių kaip atsiliepimų apie prekes, klasifikavimo tikslumą, tačiau šios reikšmės nėra reikšmingos lyginant su gautais kitų klasifikavimo metodų, naudojant tik unigramą, rezultatais.
3. Logistinės regresijos (angl. *Logistic Regression*) metodas su siūlomomis *n*-gramų (unigrama, bigrama, trigrama) junginiais pagal vidutinį daugiaklasių tekstinių duomenų klasifikavimo tikslumą 12–15 % pranoksta kitus klasifikavimo metodus, tokius kaip Bajesas (angl. *Naïve Bayes*), atsitiktinis miškas (angl. *Random Forest*), sprendimų medis (angl. *Decision Tree*), atraminių vektorių mašinų (angl. *Support Vector Machine*) ir daugiasluoksnis perceptronas (*angl. Multilayer Perceptron*), klasifikuojant didelės apimties daugiaklasius trumpų tekstinių žinučių duomenis, kai taikomos natūralios kalbos apdorojimo, žodžių dažnių ir kitos triukšmo mažinimo savybės – žodžių segmentavimas, nereikšmingų žodžių šalinimas, didžiųjų raidžių keitimas mažosiomis, žodžių normalizavimas.
4. Siūlomas metodas, kuriame apjungiama logistinė regresija (angl. *Logistic Regression*), *n*-gramos (unigramos, bigramos ir trigramos) junginiai ir žodžių dažniai, leidžia pasiekti didesnį klasifikavimo tikslumą (vidutiniškai 57–58 %) vykdant trumpų tekstinių žinučių, pavyzdžiui, atsiliepimų apie prekes, klasifikavimo užduotis.

INVESTIGATION OF MULTI-CLASS CLASSIFICATION METHODS FOR TEXTUAL DATA

Research context

A largely scalable and distributable computing environment provides a possibility of carrying out various data-intensive, natural language processing, and machine-learning tasks. One of them is a multi-class text classification into predefined classes, with issues involving text classification, recently investigated by many data scientists. Text classification into predefined classes is basically recognized as a sentiment analysis that analyzes the emotional tone for the given content and by the classification task assigns the meaning of a sentiment, e.g. either positive or negative. Text classification interrelates with a variety of elements and subjects which renders technical possibilities of big textual data classification, involving mathematical, statistical, data engineering, pattern recognition, machine learning, modeling, high-performance computing, and natural language processing methods and techniques. Otherwise, this is nothing else but only new forms of data analysis including knowledge gathering by using in-memory computing and computer network possibilities. It involves an integral part of intelligence and new emerging fields that contain collection and analysis of natural language data by delivering solutions for decision makers. The focus of the investigation is on comparing multi-class classifiers by evaluating the text classification accuracy, based on the size of training data, the number of *n*-grams, tokens, and other modern methods, such as a part of speech, term frequency, etc. In experiments, product-review data from Amazon[4] are analyzed. Particularly, such a product-review collects useful information that might help each potential customer to decide whether to order this product or service, or not. On the other hand, some product-reviews are useless, i.e. they do not provide significant information and have a negative rating, because the delivery happened one day later than expected. Also, there are other types of product-reviews that provide neither useful nor useless information about the product or service.

This is a challenging issue to deal with all the data and informatics engineering must prepare the tools and modern methods, based on how to find the effective and accurate ways to work with such challenges. The thesis investigates the impact of cloud computing technology on the classification and modern natural language processing methods. The research and experiments are implemented in Apache Spark, i.e. the in-memory intensive computing platform managed in the cloud computing environment. The aim of this dissertation is to propose a combination of data feature selection, and classifiers that determines a multi-class classification problem with a higher classification accuracy for large-scale short text product-review data.

[4] Amazon is a registered trademark. More: https://amazon.com

Statement of the problem

Text is a valuable source of information when it comes to knowledge gathering about online and purchase behavior or emotional influence effect on what to buy. Usually product-reviews influence the role of emotion and decision of a consumer whether to buy this product or service on not. Precisely, the attention of internet-based retailers is always focused on the tools and methods how to promote their products and increase the sales generated revenue and, at the same time, to collect the customers' feedback. Product-reviews (or online-reviews) usually consist of complex large-scale textual data. This type of data is used for buying or selling online, e.g. in specialized online data stores, and other specific internet directories. However, some product-reviews are more helpful and influential to the customers and sellers than others, but they are not always very evident because of the big data impact. Product-reviews might be also professional, unprofessional, short or long, with psychological and behavioral or perception and judgment aspects, which indicates distinctive personality types. There are several papers that analyze the classification problem of product-reviews, but the authors have not proposed a multi-class classification method that determines a higher classification accuracy of large-scale textual data, using data feature selection with a combination of *n*-grams. Comparison studies usually include Naïve Bayes and Support Vector Machine, but not Logistic Regression as a comparable classification method. Also, some authors have shown that changing parameters of classification methods have a lower impact on the classification accuracy than reasonably preparing the text corpora, by applying natural language processing and data feature selection techniques.

Therefore, the goal of this dissertation is to propose a combination of data feature selection and classifiers that determine the product-review classification problem with a higher classification accuracy for large-scale product-review data. In other words, the idea is to identify and accurately assign prediction of a class to unknown product-reviews, when a training set of review data with class labels is given. New in-memory computing technology evolution capabilities open the doors for innovative ways of processing and classifying large-scale natural language data. This scientific investigation is about the impact changes in large-scale natural language data, development of in-memory data analytics frameworks, machine learning, and multi-class classification methods.

Research object

The research object is the natural language processing methods for multi-class classification, based on modern cloud computing technology solutions and data analytics frameworks.

Research aim and objectives

The aim is to propose a combination of data feature selection and classifiers that determine the multi-class classification problem with a higher classification accuracy for large-scale short text product-review data. To accomplish the aim of research, the following tasks were performed:

1. To investigate data intensive technologies, multi-class classification, and natural language processing methods and techniques.
2. To compare data intensive technologies, multi-class classification methods, and data feature selection for large-scale textual data, by performing a comparative analysis of the realized and investigated methods, using real data sets, based on multi-class classification performance criteria: classification accuracy, precision, recall, error rate, F1-measurement.
3. To propose data features selection that improve the multi-class classification accuracy with the given data, e.g. for short texts such as product-review messages.
4. To propose a modified workflow model, including the corresponding methods and techniques for multi-class classification, suitable to classify short-text messages more accurately.

Research methods

The whole research methodology, applied in this thesis, is mainly based on:

1. Bibliographic research of the stated research questions and objectives was used and helped to identify, select, and evaluate the research evidence relevant to these questions and objectives.
2. The analysis of the scientific, experimental and practical achievements in the fields of machine learning with textual data classification in the cloud computing technology, the use of information retrieval, organization, analysis, benchmarking and aggregation methods.
3. Quantitative and qualitative information gathering was used in the problem-solving procedures, e.g. collection of experimental data for multi-class classification methods.
4. Constructive research procedure for producing new constructions, found to offer the solution to the real-world challenges, and to make some contribution to the theory of the discipline in which it can be applied.
5. A case-based and controlled experiments were used in the experimental part in this thesis. The experimental research methodology is described in Chapter 3.
6. Software development methods were used in the experimentation phase for constructing multi-class classification methods, and data feature selection techniques for large-scale textual data, based on multi-class classification performance criteria: classification accuracy, precision, recall, error rate, F1-measurement.

Scientific contribution of the research

The following scientific contribution is presented in the dissertation:

1. The multi-class classification methods for large-scale short text product-review data are experimentally investigated.
2. The proposed combination of *n*-grams (unigram, bigram, trigram) is effective in selecting term frequency data features, applied in Logistic Regression, Support Vector Machine, Naïve Bayes, Random Forest, Decision Tree, Multi-Layer Perceptron classifiers, and determines the multi-class classification problem with a higher classification accuracy.

3. The presented methodology for multi-class text classification tasks using the resampling of data feature selection (noise reduction, bags of words and term frequency) are suitable for datasets that contain a large amount of short texts such as product-review messages. This type of data usually is used for buying or selling online, e.g. in specialized online data stores, internet-based retailers, telecommunication, and specific internet directories.
4. Comparative analysis of cloud-based big data analytics frameworks has shown that Apache Spark analytics framework, used in the cloud computing technology, is suitable to scale the amount of textual data and apply a variety of classifications using machine learning algorithms in a horizontally distributable computer network.

Practical value of the research

The results of presented methodology for multi-class text classification, and the proposed feature selection method can be used in a variety of large-scale textual data processing systems and tools:

1. To deal with large-scale data, select and classify negative and positive or neutral product-reviews, and promote the most accurate, positive ones that will allow us to increase the incomes when selling various products and services, to provide additional sale services, or to support a customer retention program, e.g. to detect unsatisfied customers.
2. To find the optimal structures and their values, to implement the algorithms, understand and predict the textual data, to support decision making and the knowledge gathering process in science or business.
3. To investigate large network datasets for market research, detecting antisocial online behavior, classifying text documents that are related to cybersecurity area, e.g. malicious domains, source code vulnerability, phishing identification, etc.

Statements to be defended

The following statements to be defended in the dissertation are presented:

1. Big data analytics frameworks can be successfully used in the in-memory intensive operations for machine learning, e.g. classification algorithms.
2. The proposed method of data feature selection, based on a combination of *n*-grams (unigram, bigram, trigram), term frequency allows us to achieve a higher classification accuracy with short texts such as product-review messages, when the method is used with Logistic Regression, Support Vector Machine, Naïve Bayes, Random Forest, Decision Tree, Multi-Layer Perceptron.
3. The Logistic Regression method that has outperformed Support Vector Machine, Naïve Bayes, Random Forest, Decision Tree, and Multi-Layer Perceptron classification methods with the given large-scale multi-class textual data sets that contains a proposed combination of *n*-grams (unigram, bigram, trigram) as compared to unigrams, and applied to term frequency, and noise reduction techniques such as tokenization, stop word removal, lowercasing, and stemming.

4. The proposed multi-class classification method with a combination of *n*-grams (unigram, bigram, trigram) achieves a higher classification accuracy using short texts such as product-review data.

Publikacijų sąrašas

Straipsniuose recenzuojamuose moksliniuose žurnaluose:

1. Pranckevičius T. Parallel data processing services based on Cloud computing technology. *Information Sciences,* 73: 64–73, 2015. ISSN 1392-0561.

2. Pranckevičius T., Marcinkevičius V. Comparison of Naïve Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification. *Baltic J. Modern Computing*, Vol. 5, No. 2, 221–232, 2017. ISSN 2255-8950.

Straipsniuose recenzuojamuose konferencijų leidiniuose:

3. Pranckevičius T., Marcinkevičius V. Logistic Regression and Tokenization Methods Applied for Multi-Class Text Classification. *2016 IEEE 4th Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE)*, Vilnius, 2016. ISBN: 978-1-5090-4473-3.

Straipsniai kitose konferencijų suvestinėse:

4. Pranckevičius T., Marcinkevičius V. Classification and visualization algorithms on cloud computing: issues and advantages. *Data analysis methods for software systems: 7th international workshop*, Druskininkai, December 3–5, 2015: [abstract book] Vilnius: Vilniaus universiteto Matematikos ir informatikos institutas, 2015. ISBN 978-9986-680-58-1.

Informacija apie disertacijos autorių

Tomas Pranckevičius gimė 1983 m. rugsėjo 6 d. Panevėžyje. 2006 m. Vilniaus universitete įgijo vadybos ir verslo administravimo bakalauro laipsnį. 2009 m. Vilniaus Gedimino technikos universitete įgijo vadybos ir verslo administravimo magistro laipsnį. 2009 m. studijavo Freibergo kalnakasybos ir technologijų universitete (Vokietija). 2011 m. Kauno Technologijos universitete įgijo informatikos inžinerijos magistro laipsnį. 2011–2016 m. studijavo doktorantūroje (technologijos mokslai, informatikos inžinerija) Vilniaus universitete, Matematikos ir informatikos institute.

Tomas Pranckevičius

DAUGIAKLASIŲ TEKSTINIŲ DUOMENŲ KLASIFIKAVIMO METODŲ TYRIMAS

Daktaro disertacijos santrauka
Technologijos mokslai, informatikos inžinerija (07 T)
Redaktorė Jorūnė Rimeisytė

Tomas Pranckevičius

INVESTIGATION OF MULTI-CLASS CLASSIFICATION METHODS FOR TEXTUAL DATA

Summary of Doctoral Dissertation
Technological Sciences, Informatics Engineering (07 T)
Editor Janina Kazlauskaitė