VILNIAUS UNIVERSITETAS

Konstantinas

KOROVKINAS

Hibridinis tekstinių duomenų metodas nuomonių analizei

DAKTARO DISERTACIJOS SANTRAUKA

Gamtos mokslai
Informatika N 009

VILNIUS 2020

Disertacija rengta 2015–2020 metais Vilniaus universitete.

Mokslinis vadovas:
prof. dr. Gintautas Garšva (Vilniaus universitetas, gamtos mokslai, informatika – N 009).

Gynimo taryba:

Pirmininkė – prof. dr. Olga Kurasova (Vilniaus universitetas, gamtos mokslai, informatika – N 009).

Nariai:
prof. dr. Robertas Damaševičius (Kauno technologijos universitetas, technologijos mokslai, informatikos inžinerija – T 007),
prof. habil. dr. Gintautas Dzemyda (Vilniaus universitetas, gamtos mokslai, informatika – N 009),
prof. habil. dr. Janusz Kacprzyk (Lenkijos mokslų akademijos Sistemų tyrimų institutas, gamtos mokslai, informatika – N 009),
dr. Virginijus Marcinkevičius (Vilniaus universitetas, gamtos mokslai, informatika – N 009).

Disertacija bus ginama viešame Gynimo tarybos posėdyje 2020 m. rugsėjo mėn. 23 d. 12:00 val. Vilniaus universiteto Duomenų mokslo ir skaitmeninių technologijų instituto 203 auditorijoje. Adresas: Akademijos g. 4, LT-08412, Vilnius, Lietuva.
Disertacijos santrauka išsiuntinėta 2020 m. rugpjūčio mėn. 21 d.
Disertaciją galima peržiūrėti Vilniaus universiteto bibliotekoje ir VU interneto svetainėje adresu: `https://www.vu.lt/naujienos/ivykiu-kalendorius`

VILNIUS UNIVERSITY

Konstantinas

KOROVKINAS

Hybrid Method for Textual Data Sentiment Analysis

SUMMARY OF DOCTORAL DISSERTATION

Natural Sciences
Informatics N 009

VILNIUS 2020

This dissertation was written between 2015–2020 at Vilnius University.

Scientific supervisor:
Prof. Dr. Gintautas Garšva (Vilnius University, Natural Sciences, Informatics – N 009).

This doctoral dissertation will be defended in a public meeting of the Dissertation Defence Panel:

Chairman – Prof. Dr. Olga Kurasova (Vilnius University, Natural Sciences, Informatics – N 009).

Members:
Prof. Dr. Robertas Damaševičius (Kaunas University of Technology, Technological Sciences, Informatics Engineering – T 007),
Prof. Habil. Dr. Gintautas Dzemyda (Vilnius University, Natural Sciences, Informatics – N 009),
Prof. Habil. Dr. Janusz Kacprzyk (Polish Academy of Sciences, Systems Research Institute, Natural Sciences, Informatics – N 009),
Dr. Virginijus Marcinkevičius (Vilnius University, Natural Sciences, Informatics – N 009).

The dissertation shall be defended at a public meeting of the Dissertation Defence Panel at 12:00 p.m. on 23^{th} of September 2020 in Room 203 of the Institute of Data Science and Digital Technologies of Vilnius University.
Address: Akademijos street 4, LT-08412, Vilnius, Lithuania.
The summary of the doctoral dissertation was distributed on the 21^{th} of August, 2020.
The text of this dissertation can be accessed at the library of Vilnius University, as well as on the website of Vilnius University: `https://www.vu.lt/naujienos/ivykiu-kalendorius`

SANTRAUKA

1. ĮVADAS

Tyrimo kontekstas

Tekstinių duomenų analizė tapo labai populiari, kai žmonės pradėjo naudotis internetu, tiksliau tariant – kai atsirado el. parduotuvės ir socialiniai tinklai, tinklaraščiai ir kitos platformos, kuriose žmonės gali rašyti komentarus. Šiais laikais, norint gauti nuomonę apie apklausas, socialinius, ekonominius ar kitus įvykius, visą informaciją galima rasti internete. Pagrindinis tyrimų, susijusių su sentimentų analize, tikslas – gauti autorių jausmus, išreikštus teigiamais ar neigiamais komentarais. Ši analizė atliekama keliais lygiais: dokumento, sakinio ir aspekto. Anot autorių Pang ir Lee, terminas „sentimentas“ buvo pavartotas 2001 m. ir vėliau 2002 m. Nuomonių gavyba – kitas terminas, atitinkantis sentimentų analizę ir pavartotas 2003 m. autorių Dave ir kt. straipsnyje. Autoriai apibūdino idealų nuomonių gavybos įrankį, kuris, anot jų, „apdorotų pateikto elemento paieškos rezultatų rinkinį, sudarydamas produktų atributų (kokybė, savybės ir kt.) sąrašą ir apibendrindamas nuomones apie kiekvieną iš jų (prasta, mišri, gera)“. Autorius Liu pateikė sentimentų analizės apibrėžimą – „studijų sritis, analizuojanti žmonių nuomones, jausmus, vertinimus, požiūrius ir emocijas, susijusias su subjektais ir jų požymiais, išreikštais rašytiniu tekstu“. Anot jų, ši sritis reprezentuoja didžiulę probleminę erdvę dėl „daug susijusių pavadinimų ir šiek tiek skirtingų užduočių“. Anot Liu, „sentimentų analizė, nuomonių gavyba, nuomonių analizė, sentimentų gavyba, subjektyvumo analizė, įtakos analizė, emocijų analizė ir apžvalgų gavyba“ patenka į sentimentų analizės apibrėžimo sritį. Iš esmės sentimentų analizė yra dalijama į leksikonu grįstus metodus ir mašininio mokymosi (ML) metodus. Tyrėjai dažnai sujungia šiuos du metodus, siekdami gauti geresnių rezultatų. Leksikonu grįsti metodų tyrimai apima dokumentų orientacijos apskaičiavi-

mą pagal semantinę žodžių orientaciją ar frazę iš dokumento. Šiuo metu sentimentų analizė smarkiai išpopuliarėjo ir vis dar populiarėja dėl tekstinių duomenų kiekio, prieinamo internetu, kuris gali būti labai naudingas įmonėms, naudotojams ir pan. Leksikonu grįstų metodų naudojimas su dideliais duomenų kiekiais, gautais iš socialinių tinklų, nėra labai efektyvus dėl nestruktūrizuoto formato ir teksto ypatumų, neformalios ir dinamiškos kalbos pobūdžio, naujo slengo, santrumpų ir naujų išraiškų. Be to, šis metodas gali žymiai padidinti skaičiavimo resursus. Taigi šioje disertacijoje tiriama nuomonių klasifikacija, naudojant mašininio mokymosi metodus.

Tyrimo problema

Nuomonių analizė yra iššūkių pilna sritis: nors šioje srityje yra padaryta daug darbų, tikslumas vis dar yra gana vidutinis dėl komentarų, žargonų, naudojamų jausmaženklių ir pan. Reikia suprasti visą sakinio esmę, kadangi net vienas žodis gali pakeisti jo poliškumą ir tai gali turėti rimtų pasekmių ypač šiuo metu aktualiose srityse tokiose kaip medicina, verslas ir t. t. Kita problema, su kuria susiduriama nuomonių analizės srityje – dideli duomenų kiekiai. Jei naudotojas nori susidaryti nuomonę apie prekę, viešbutį, skrydį ir pan. – tai gali tapti rimtu iššūkiu. Didelių duomenų kiekių atveju mašininio mokymosi algoritmų mokymosi greitis mažėja priklausomai nuo duomenų kiekio – kuo didesnis ypatybių kiekis, tuo ilgesnis skaičiavimo laikas jam reikalingas. Norint išspręsti šią problemą, tyrimuose naudojamos įvairios technikos: lygiagretinimas, grafikos procesorius, debesų kompiuterijos technologijos, naudojant vien reprezentacinių mokymosi duomenų rinkinius ir pan. Tai lemia kitą problemą – specialios aparatinės įrangos ar galingesnio kompiuterio, debesijos paslaugų tiekėjo su reikiama programine įranga ir pan. poreikį. Šioje disertacijoje nagrinėjama didelių skaičiavimo resursų problema, su kuria susiduria mašininio mokymosi algoritmai, vykdomi su didelių tekstinių duomenų kiekiu. Todėl didžiausias dėmesys

skiriamas mašininio mokymosi algoritmų vykdymo greičiui didinti, neprarandant ar šiek tiek prarandant klasifikavimo tikslumą, be specialios programinės ar aparatinės įrangos poreikio.

Tyrimo objektas

Pagrindinis šio tyrimo objektas yra tekstinių duomenų klasifikavimo metodai, jų vykdymo greitis ir tikslumas didelės apimties duomenų rinkiniuose, nuomonių analizė.

Darbo tikslas ir uždaviniai

Tyrimo tikslas – pasiūlyti hibridinį nuomonių analizės metodą su rekomenduojamų parametrų rinkiniu dideliems tekstiniams duomenims, kurio vykdymo greitis būtų spartesnis, o klasifikavimo tikslumas – panašus, lyginant su klasikiniais metodais.

Tyrimo uždaviniai:

1. Pasiūlyti hibridinį metodą, kuris padidintų pasirinktų klasikinių mašininio mokymosi algoritmų, kurie dažniausiai naudojami tekstinių duomenų nuomonių analizei, klasifikavimo greitį ir išlaikytų panašų klasifikavimo tikslumą.
2. Atlikti siūlomo metodo eksperimentinį vertinimą ir parinkti jam rekomenduojamą parametrų rinkinį bei pagerinti klasifikavimo tikslumą.
3. Atlikti siūlomo hibridinio metodo palyginimą su kitų autorių darbais, naudojant didelės apimties tekstinių duomenų rinkinius, ir įvertinti gautus rezultatus.

Tyrimo metodai ir priemonės

Disertacijoje taikomi šie metodai:

1. Bibliografiniai nuomonių analizės srities tyrimai padėjo suformuluoti tyrimo uždavinius ir tikslus.
2. Susijusių darbų analizė padėjo pasirinkti siūlomo hibridinio metodo mašininio mokymosi algoritmus.
3. Siūlomas hibridinis metodas tekstinių duomenų sentimentų analizei aprašytas skyriuje „Tyrimo metodika“.

4. Eksperimentinių tyrimų metodika, taip pat kaip eksperimentiniai tyrimai siūlomo metodo lyginamai analizei, yra aprašyti skyriuje „Eksperimentai ir rezultatai“.
5. Išvados formuluojamos po kiekvieno skyriaus, o tyrimo pabaigoje pateikiamos bendrosios išvados.

Siūlomam hibridiniam metodui kurti ir eksperimentiniams tyrimams atlikti buvo naudojami: Python programavimo kalba, scikit-learn[1] – mašininio mokymosi biblioteka. Disertacijai rengti naudotas LaTeX[2], dokumentų kūrimo įrankis. Grafiniams rezultatams ir schemoms pateikti – latex Tikz[3] paketas.

Mokslinis naujumas

Disertacijoje pasiūlytas hibridinis tekstinių duomenų metodas nuomonių analizei su rekomenduojamu parametrų rinkiniu, pritaikytu didelės apimties duomenų rinkiniams. Metodą sudaro:

1. SpeedUP metodas – tai yra pagrindinė pasiūlyto metodo dalis, kurios tikslas yra padidinti klasikinių mašininio mokymosi algoritmų klasifikavimo greitį.
2. k-Means klasterizavimo metodas – šis metodas parenka mokymosi duomenis.
3. PSO derinimo metodas – šis metodas atlieka tiesinės atraminių vektorių mašinos (LSVM) hiperparametrų derinimą.
4. Ansamblis – tai yra paskutinė siūlomo hibridinio metodo dalis, kuri vykdo mašininio mokymosi algoritmų jungimą ir balsavimą.

Šioje disertacijoje buvo apžvelgti mašininio mokymosi algoritmai tekstinių duomenų nuomonių analizei ir išrinkti penki labiausiai paplitę. Pasirinkti mašininio mokymosi algoritmai buvo vertinami pagal šiuos matus: efektyvumo rodikliai, vidutinis reitingas ir statistinis reikšmingumas.

[1] `https://scikit-learn.org/`

[2] `https://www.latex-project.org/`

[3] `https://www.overleaf.com/learn/latex/TikZ_package`

Eksperimentiškai parinktas ir pateiktas rekomenduojamų parametrų rinkinys pasiūlytam hibridiniam metodui. Rezultatai parodė, kad pasiūlytas metodas padidino klasikinių mašininio mokymosi algoritmų klasifikavimo greitį didelės apimties tekstiniuose duomenų rinkiniuose, šiek tiek prarandant tikslumą; jis taip pat gali konkuruoti su naujausiais metodais.

Skirtingai nuo anksčiau pasiūlytų metodų, SpeedUP metodas automatiškai vykdo visas pasiūlyto hibridinio metodo dalis, priklausomai nuo nurodytų parametrų, kurie rekomenduojami šiame tyrime bei nustatyti kaip numatytieji SpeedUP metode. Atsižvelgiant į nustatytą poaibio dydį yra apskaičiuojamas mokymosi duomenų kiekis ir priklausomai nuo jo k-Means metodas parenka atitinkamą kiekį mokymosi duomenų (ansamblio metodo atveju parenka tiek mokymosi duomenų rinkinių, kiek yra pasirinktų klasifikatorių ansambliui) ir perduoda jį į LSVM įvestį; jei k-Means metodas yra išjungtas, duomenų rinkiniai bus parinkti atsitiktiniu būdu. PSO metodas automatiškai parenka LSVM baudos parametrą C ir jo mokymas atliekamas su šiuo parametru; ansamblio atveju parenkamas baudos parametrų kiekis, atitinkantis klasifikatorių skaičių. Visi skaičiavimai, mokymosi duomenų parinkimas, testavimo duomenų dalijimas į poaibius, parametrų derinimas, mašininio mokymosi algoritmų jungimas į ansamblį, rezultatų jungimas ir balsavimas atliekami automatiškai SpeedUP metode.

Pasiūlytas hibridinis metodas gali būti taikomas klasifikuojant nuomones tekstiniuose duomenyse bei pritaikomas naudoti su dideliais duomenų kiekiais nenaudojant superkompiuterių.

Praktinė vertė

Tekstinių duomenų nuomonių analizė vis dar labai sudėtinga sritis, tačiau labai plačiai taikoma produktų apžvalgoms, klientų skaičiaus prognozei, sukčiavimui aptikti, rinkimams ir pan., pasiūlytas metodas gali būti:

1. Sėkmingai pritaikytas pirmiau minėtose srityse kuriant naujus modelius arba tobulinant esamus.
2. Pritaikytas darbui su dideliais tekstinių duomenų rinkiniais ir klasifikuoti nuomonėms nenaudojant didelio našumo kompiuterių.
3. Naudingas atliekant apklausas, nes nuomonės klasifikavimui gali būti imamos iš socialinių tinklų, straipsnių komentarų ir pan. Tai leidžia susidaryti kitokią nuomonę dominančia tema.

Ginamieji teiginiai

1. Pasiūlytas hibridinis tekstinių duomenų metodas nuomonių analizei, pritaikytas didelės apimties duomenų kiekiams, gali padidinti klasikinių mašininio mokymosi algoritmų, tokių kaip – tiesinė atraminių vektorių mašina, logistinė regresija (LR), sprendimų medis (DT) ir atsitiktinis miškas (RF) – klasifikavimo greitį, o klasifikavimo tikslumo nuostoliai nėra labai dideli.
2. Geriausi rezultatai pasiekti, kai pasiūlytas hibridinis metodas su rekomenduojamu parametrų rinkiniu naudojamas su LSVM, lyginant su tuo, kai jis naudojamas su daugialypiu naiviuoju Bajesu (MNB), logistine regresija, sprendimu medžiu ir atsitiktiniu mišku.
3. Pasiūlytas hibridinis metodas galėtų būti naudojamas kaip alternatyvus metodas klasikiniams ir moderniems metodams, darbui su dideliais tekstinių duomenų rinkiniais. Be to, PSO parametrų derinimo metodas galėtų konkuruoti su tokiais populiariais metodais kaip atsitiktinė paieška (RS) ir Bajeso optimizavimas (Bopt).

Rezultatų aprobavimas ir publikavimas

Tyrimo rezultatai pristatyti trijose tarptautinėse konferencijose ir dviejose konferencijose Lietuvoje. Disertacijos tema paskelbti trys straipsniai recenzuojamuose žurnaluose (1 žurnalas yra įtrauktas į ISI Web of Science duomenų bazę), keturi straipsniai

– periodiniuose konferencijų leidiniuose, 1 santrauka konferencijų santraukų leidinyje.

Disertacijos struktūra

Disertaciją sudaro 4 skyriai, bendrosios išvados, literatūros sąrašas ir priedai. Disertacijos apimtis – 170 puslapių, 55 lentelės ir 33 paveikslai. Literatūros sąrašą sudaro 212 įvairių šaltinių, įskaitant knygas, mokslinius straipsnius, patentus, technines ataskaitas ir interneto šaltinius.

2. TYRIMO METODIKA

Šioje disertacijos dalyje pateiktas hibridinis tekstinių duomenų metodas nuomonių analizei. Metodo tikslas – padidinti klasikinių mašininio mokymosi algoritmų klasifikavimo greitį išlaikant panašų į šių algoritmų klasifikavimo tikslumą. Taip pat pateikiami naudojamų duomenų bazių aprašymai ir metodo vertinimo matai.

Pasiūlytas hibridinis metodas

Pasiūlytą hibridinį metodą sudaro keturios dalys: klasikinis mašininio mokymosi algoritmas, k-Means klasterizavimas, dalelių spiečiaus optimizavimo metaeuristika ir ansamblis. Šie metodai yra integruoti į SpeedUP metodą. SpeedUP metodo išraiška (numatytieji parametrai nustatyti atsižvelgiant į eksperimentinius rezultatus):

$$\textbf{SpeedUP}(ml, kmeans, ensemble, pso, Subset_{size}, D_{text}, num_{class}) \tag{2.1}$$

Parametrai:

ml – simbolių eilutė (angl. *string*), ‘LSVM’, ‘MNB’, ‘RF’, ‘DT’ arba ‘LR’ (numatytasis = ‘LSVM’); *ml* nurodo klasikinį mašininio mokymosi algoritmą, kuris bus naudojamas.

kmeans – sveikasis skaičius (angl. *integer*), 0 arba 1 (numatytasis = 1); mokymosi duomenų parinkimo metodas. Jei *kmeans* = 1, tai k-Means klasterizavimo metodas bus

naudojamas mokymosi duomenų aibei parinkti, kai ***kmeans*** = 0 – mokymosi duomenų aibė bus parenkama naudojant atsitiktinį parinkimą (angl. *random sampling*).

ensemble – sveikasis skaičius (angl. *integer*) (numatytasis = 0); ***ensemble*** parametras nurodo balsuotojų (angl. *voters*) skaičių metode. Jei ***ensemble*** = 0 arba 1 (rekomenduojama, kai *pso* = 1), tai ansamblio metodas nebus naudojamas.

pso – sveikasis skaičius (angl. *integer*), 0 arba 1 (numatytasis = 1); *pso* = 1 rodo, kad parametrams derinti bus naudojama dalelių spečiaus optimizavimo metaeuristika (tik 'LR' arba 'LSVM'). 'MNB', 'RF' ir 'DT' atveju – *pso* = 0.

$\mathbf{Subset_{size}}$ – sveikasis skaičius (angl. *integer*) (numatytasis = 30 000); šis parametras nurodo poaibių dydį, į kurį bus dalijama testavimo duomenų aibė. Kai $Subset_{size}$ = 30 000, testavimo duomenų aibė yra dalijama į lygius poaibius, kurių dydis – 30 000 įrašų.

$\mathbf{D_{text}}$ – tekstinių duomenų aibė.

$\mathbf{num_{class}}$ – sveikasis skaičius (angl. *integer*) (numatytasis = 2); šis parametras nurodo skirtingų klasių (poliariškumo) skaičių duomenų rinkinyje.

Pasiūlyto hibridinio metodo diagrama pavaizduota 2.1 paveiksle. Raudonai apibrėžti elementai žymi hibridinio metodo dalis.

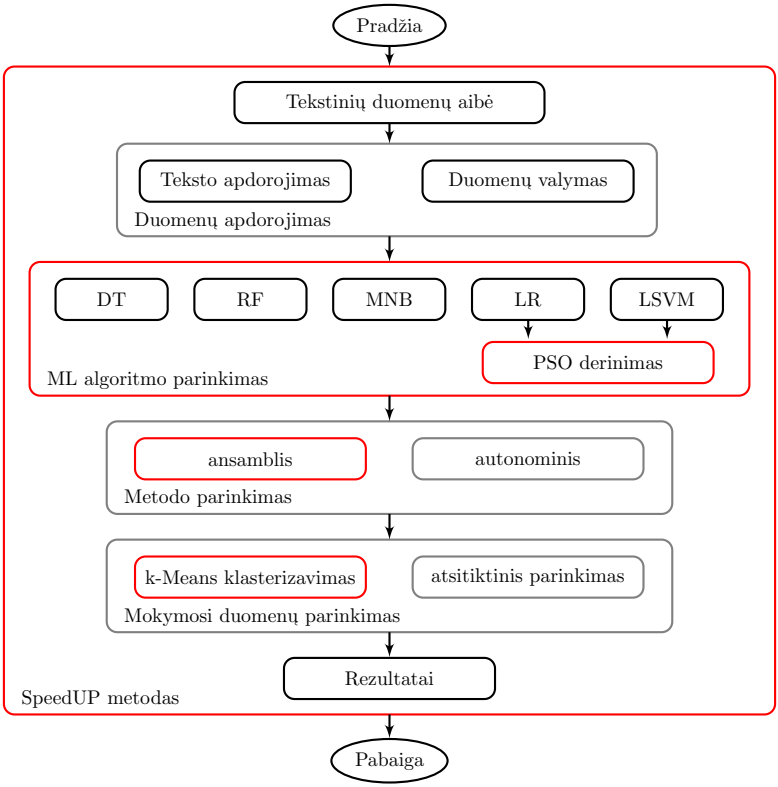
Diagramos žingsniai:

- **SpeedUP metodas**. Tai pagrindinė hibridinio metodo dalis; kitos dalys yra integruotos į jį. Šio metodo tikslas yra pagreitinti klasikinių mašininio mokymosi algoritmų klasifikavimą.
- **Tekstinių duomenų aibė**. Tekstinių duomenų aibė, kuri bus naudojama nuomonių analizei, nuskaitoma iš duomenų šaltinio ir apdorojama.
- **Duomenų apdorojimas**. Šis žingsnis apima du veiksmus: teksto apdorojimą ir duomenų valymą. Teksto apdorojimas apima tokius veiksmus kaip konvertavimas į mažąsias raides, nereikalingų leksemų šalinimas: grotelės, @, skaičiai, „http“

nuorodos, skyrybos ženklai, naudotojų vardai ir pan. Duomenų valymo veiksmas tikrina, ar yra tuščių eilučių duomenų rinkinyje ir jei yra – pašalina. Šio žingsnio tikslas yra paruošti duomenis SpeedUP metodui, kuriame jie bus paruošti mašininio mokymosi algoritmams apdoroti.

- **ML parinkimas**. Šiame žingsnyje parenkamas mašininio mokymosi algoritmas, kuris bus naudojamas SpeedUP metode. Parinkimas atliekamas keičiant *ml* reikšmę SpeedUP metode (žr. 2.1 išraišką). Šis žingsnis taip pat apima PSO parametrų derinimo metodą, kuris yra naudojamas su logistine regresija ir tiesine atraminių vektorių mašina, siekiant padidinti klasifikavimo tikslumą. Parinkimas atliekamas nustatant reikšmę $pso = 1$ SpeedUP metode (žr. 2.1 išraišką).
- **Metodo parinkimas**. Šiame žingsnyje nustatoma, ar bus naudojamas autonominis mašininio mokymosi algoritmas ar ansamblis. Ansamblis naudojamas, kai SpeedUP metode (žr. 2.1 išraišką) rekšmė $ensemble > 1$. Ši reikšmė nustato, kiek mašininio mokymosi algoritmų sudarys ansamblį. Svarbu paminėti, kad ansamblį sudaro tik vienos rūšies mašininio mokymosi algoritmai. Reikšmė *ml* SpeedUP metode nustato, kuris ML algoritmas sudarys ansamblį; jei $ml =$ 'LSVM', tai ansamblį sudarys tik LSVM algoritmai.
- **Mokymosi duomenų parinkimas**. Šiame žingsnyje parenkamas mokymosi duomenų parinkimo metodas. Yra galimybė atsitiktinai parinkti mokymosi duomenis arba pritaikius k-Means klasterizavimo metodą. Tai atliekama nustatant parametro *kmeans* reikšmę SpeedUP metode (žr. 2.1 išraišką).
- **Rezultatai**. Nuomonių klasifikavimo rezultatai pateikiami statistinėmis priemonėmis: tikslumas (angl. *Accuracy*; ACC), preciziškumas (angl. *Precision*; PPV – teigiama prognozinė vertė, NPV – neigiama prognozinė vertė), prisiminimas (angl. *Recall*; TPR – tikroji teigiama norma, TNR – tikroji neigiama norma) ir F1 rodiklis ($F_1 score$). Modelio

prognozinei kokybei įvertinti naudojamas plotas po sprendimus priimančiojo ypatybių kreive (angl. *Area under the receiver operating characteristics*; (AUC)).



2.1 paveikslas: Pasiūlytas hibridinis metodas

Duomenų aibės

Eksperimentams ir rezultatams palyginti pasiūlytas metodas yra testuojamas su didžiausiomis sužymėtomis atvirai prieinamomis tekstinėmis duomenų aibėmis:

- Stanfordo Twitter nuomonių korpuso duomenų aibė (angl.

Stanford Twitter sentiment corpus dataset[4]; sentiment140).

- Amazon klientų apžvalgų duomenų aibė (angl. *Amazon customer reviews dataset*[5]; AmazonTest). Ši duomenų aibė yra sudaryta iš duomenų aibės, pristatytos autorių Zhang ir kt., kuri originaliai sudaryta iš Amazon produktų duomenų aibės (angl. *Amazon product data dataset*[6]), atsitiktinai parenkant mokymosi ir testavimo duomenis iš kiekvienos apžvalgos klasės, įvertintos nuo 1 iki 5 balų. Vėliau gauta duomenų aibė buvo rekonstruota į dviejų klasių duomenų aibę.
- Amazon produktų duomenų aibė[6] pristatyta autorių McAuley ir kt. Konkrečiau iš Amazon produktų duomenų aibės pasirinktos šios duomenų aibės: knygos (*Books*), elektronika (*Electronics*), Kindle parduotuvė (*KindleStore*), mobiliojo ryšio telefonai ir jų priedai (*Phones&Accessories*).
- Metodo praktiniam pritaikymui su realiais duomenimis: atsiliepimų apie konkretų asmenį duomenų aibė (Person) ir atsiliepimų apie įvykį duomenų aibė (Event). Šios dvi aibės buvo surinktos ir sužymėtos paties autoriaus.

3. EKSPERIMENTAI IR REZULTATAI

Šioje disertacijos dalyje yra atliekamas siūlomo metodo eksperimentinis vertinimas ir rekomenduojamo parametrų rinkinio parinkimas. Pasiūlytas metodas yra lyginamas su kitų autorių darbais. Taip pat pateikiami eksperimentai su realiais duomenimis. Eksperimentai atliekami su didžiausiomis sužymėtomis atvirai prieinamomis duomenų aibėmis. Pasiūlyto metodo žymėjimas (figūriniuose skliaustuose esantys kintamieji yra neprivalomi):

$$\mathbf{ML}^{\{\mathbf{PSO}\}}_{\{\mathbf{n}\}}\mathbf{\_\{km\}\_s\_SpeedUP}$$

[4] `http://help.sentiment140.com/`

[5] `https://www.kaggle.com/bittlingmayer/amazonreviews/`

[6] `http://jmcauley.ucsd.edu/data/amazon/`

kur **ML** – mašininio mokymosi algoritmas; **n** – nurodo ML algoritmų skaičių ansamblyje, jei nenurodyta – ansamblis nenaudojamas; **PSO** – nurodo, kad įjungtas PSO metodas; **km** – nurodo, kad įjungtas k-Means metodas; **s** – nurodo poaibio dydį; **SpeedUP** – SpeedUP metodas.

3.1 Eksperimentų ciklai

Šioje disertacijoje atlikti šeši eksperimentų ciklai:

1. **Klasikinių mašininio mokymosi algoritmų eksperimentų ciklas**. Šiame cikle atliekami eksperimentai su penkiais klasikiniais mašininio mokymosi algoritmais: daugialypiu naiviuoju Bajesu, atsitiktiniu mišku, sprendimų medžiu, tiesine atraminių vektorių mašina ir logistine regresija. Šiam eksperimentų ciklui duomenų aibė dalijama naudojant standartinį dalijimą į mokymosi (70 % duomenų) ir testavimo (30 % duomenų) duomenų aibes. Šio eksperimentų ciklo tikslas yra palyginti klasikinių mašininio mokymosi algoritmų rezultatus ir nustatyti bazines reikšmes palyginimui su pasiūlytu hibridiniu metodu.
2. **SpeedUP metodo eksperimentų ciklas**. Šiame cikle eksperimentai atliekami taikant SpeedUP metodą pirmiau paminėtiems klasikiniams mašininio mokymosi algoritmams. Eksperimentai atliekami su tomis pačiomis mokymosi ir testavimo duomenų aibėmis kaip ir pastarieji. Tikslas yra palyginti klasikinių mašininio mokymosi algoritmų rezultatus su SpeedUP metodu.
3. **k-Means klasterizavimo metodo eksperimentų ciklas**. Šiame eksperimentų cikle SpeedUP metodas yra papildomas k-Means klasterizavimo metodu. Tikslas yra parodyti mokymosi duomenų parinkimo efektyvumą. Eksperimentai atliekami su tomis pačiomis mokymosi ir testavimo aibėmis kaip ir ankstesni eksperimentai.
4. **Viso hibridinio metodo eksperimentų ciklas**. Šiame eksperimentų cikle SpeedUP metodas su k-Means klasteri-

zavimu yra papildomas ansamblio metodu. LR ir LSVM atveju papildomai atliekami eksperimentai integruojant PSO derinimo metodą. Tikslas yra parinkti parametrų rinkinį pasiūlytam hibridiniam metodui.

5. **Rezultatų palyginimo eksperimentų ciklas**. Šiame eksperimentų cikle yra atliekamas palyginimas su kitų autorių darbais, naudojant pasiūlytą hibridinį metodą su tomis pačiomis duomenų aibėmis.
6. **Realių duomenų eksperimentų ciklas**. Šiame eksperimentų cikle atliekami eksperimentai su realiais duomenimis viešosios nuomonės tyrimo srityje. Rezultatai yra lyginami su atsitiktine paieška, Bajeso optimizacija ir realiais rezultatais, pateiktais dviejų visuomenės nuomonės ir rinkos tyrimų institucijomis: „Vilmorus ltd.“ ir „Baltic surveys“.

3.2 Rezultatai

Eksperimentai parodė, kad geriausi rezultatai pasiekti, kai pasiūlytas hibridinis metodas naudojamas su LSVM.

3.1 lentelėje pateikti pasiūlyto metodo naudojamo su LSVM rezultatai su įjungtomis / išjungtomis metodo dalimis bei klasikinio LSVM su PSO metodu rezultatai.

Rezultatai parodė, kad LSVM_30K_SpeedUP metodo tikslumas padidėjo 1,02 %, kai jis buvo naudojamas su įjungtais PSO derinimo bei k-Means klasterizavimo metodais naudojant sentiment140 duomenų aibę ir 1,83 % – kai buvo naudojamas ansamblyje. Skirtumas lyginant ansamblį su klasikiniu LSVM – 0,59 %. Integravus parametrų derinimo metodą į klasikinį LSVM, pastarojo tikslumas padidėjo 0,38 %. Rezultatai su AmazonTest duomenų aibe taip pat parodė padidėjusį LSVM_30K_SpeedUP tikslumą. Tikslumas padidėjo 0,86 %, kai jis buvo naudojamas su įjungtais PSO derinimo bei k-Means klasterizavimo metodais ir 1,7 % – ansamblyje. Integravus PSO metodą į klasikinį LSVM, pastarojo tikslumas padidėjo 0,64 %.

3.1 lentelė: Pasiūlyto hibridinio metodo efektyvumo metrikos

Metodas	ACC	PPV	NPV	TPR	TNR	F_1*score*	AUC
			sentiment140				
LSVM_	77,10 %	76,60 %	77,62 %	78,05 %	76,16 %	77,32 %	85,36 %
LSVM_km_	77,30 %	76,78 %	77,83 %	78,26 %	76,33 %	77,51 %	85,55 %
LSVM^{PSO}_km_	**78,12 %**	**77,64 %**	**78,62 %**	**79,99 %**	**77,24 %**	**78,31 %**	**85,95 %**
LSVM_3_km_	78,51 %	77,87 %	79,18 %	79,65 %	77,36 %	78,75 %	**86,84 %**
LSVM_3^{PSO}_km_	**78,62 %**	**77,98 %**	**79,29 %**	**79,76 %**	**77,48 %**	**78,86 %**	86,41 %
LSVM_5_km_	**78,93 %**	**78,25 %**	**79,65 %**	**80,14 %**	**77,72 %**	**79,18 %**	**87,14 %**
LSVM_5^{PSO}_km_	78,81 %	78,16 %	79,49 %	79,96 %	77,66 %	79,05 %	86,55 %
LSVM	79,52 %	78,83 %	80,24 %	80,71 %	78,32 %	79,76 %	87,59 %
LSVM^{PSO}	**79,90 %**	**79,02 %**	**80,82 %**	**81,40 %**	**78,39 %**	**80,19 %**	**87,82 %**
			AmazonTest				
LSVM_	87,59 %	87,50 %	87,68 %	87,71 %	87,47 %	87,60 %	95,04 %
LSVM_km_	87,74 %	87,75 %	87,72 %	87,71 %	87,76 %	87,73 %	95,11 %
LSVM^{PSO}_km_	**88,45 %**	**88,57 %**	**88,33 %**	**88,29 %**	**88,61 %**	**88,43 %**	**95,24 %**
LSVM_3_km_	**88,90 %**	**88,90 %**	**88,89 %**	**88,89 %**	**88,90 %**	**88,90 %**	**95,78 %**
LSVM_3^{PSO}_km_	88,88 %	89,02 %	88,74 %	88,70 %	89,06 %	88,86 %	95,49 %
LSVM_5_km_	**89,29 %**	**89,30 %**	**89,27 %**	**89,27 %**	**89,30 %**	**89,28 %**	**95,93 %**
LSVM_5^{PSO}_km_	89,03 %	89,17 %	88,89 %	88,85 %	89,21 %	89,01 %	95,56 %
LSVM	89,58 %	91,77 %	87,60 %	86,95 %	92,20 %	89,29 %	96,33 %
LSVM^{PSO}	**90,22 %**	**90,20 %**	**90,24 %**	**90,24 %**	**90,19 %**	**90,22 %**	**96,43 %**

Pabraukimas „_“ reiškia, kad pabaigoje turi būti pridėta 30K_SpeedUP.

Siūlomas parametrų rinkinys hibridiniam metodui

Disertacijos tikslas yra pasiūlyti hibridinį tekstinių duomenų metodą nuomonių analizei su rekomenduojamų parametrų rinkiniu. Remiantis gautais rezultatais, 3.1 ir 3.2 išraiškose pateikiami rekomenduojami parametrai pasiūlytam hibridiniam metodui.

Autonominiam metodui:

$$\textbf{SpeedUP}(ml = \text{'LSVM'},\ kmeans = 1,\ ensemble = 0,\ pso = 1,\ Subset_{size} = 30\,000,\ D_{text},\ num_{class} = 2) \quad (3.1)$$

Ansambliui:

$$\textbf{SpeedUP}(ml = \text{'LSVM'},\ kmeans = 1,\ ensemble = 5,\ pso = 0,\ Subset_{size} = 30\,000,\ D_{text},\ num_{class} = 2) \quad (3.2)$$

3.2 lentelėje pateikti lyginamosios analizės su kitų autorių tyrimais rezultatai, kai pasiūlytas metodas buvo naudojamas su tomis pačiomis duomenų aibėmis ir tuo pačiu klasių skaičiumi jose.

3.2 lentelė: Rezultatų palyginimas su kitų autorių tyrimais

Autoriai	Duomenų aibė	Metodas	Tikslumas
Rain C. (2013)	Books	NB	84,50 %
Shaikh ir Deshpande (2016)		NB	80,00 %
Pasiūlytas metodas		LSVM^{PSO}_30K_SpeedUP	**89,50 %**
		LSVM$_3^{PSO}$_30K_SpeedUP	**89,86 %**
Rain C. (2013)	KindleStore	NB	87,33 %
Pasiūlytas metodas		LSVM^{PSO}_30K_SpeedUP	**91,27 %**
		LSVM$_3^{PSO}$_30K_SpeedUP	**91,50 %**
Haque ir kt. (2018)	Phones&	LSVM	**93,57 %**
	Accessories	MNB	90,28 %
		SGD	91,88 %
		RF	92,72 %
		LR	88,20 %
		DT	91,45 %
Wang ir kt. (2018)		CNN	85,90 %
		CFM	83,50 %
		PFM	84,20 %
		LR-BoW	83,80 %
		SVM-BoW	83,70 %
		SVM-Poly	82,30 %
		LR-WE	76,70 %
		SVM-WE	76,70 %
		FM	78,60 %
Pasiūlytas metodas		LSVM^{PSO}_30K_SpeedUP	90,57 %
		LSVM$_3^{PSO}$_30K_SpeedUP	90,83 %
		LSVM^{PSO}	93,22 %

Haque ir kt. (2018)	Electronics	LSVM	**93,52 %**
		MNB	89,36 %
		SGD	92,61 %
		RF	92,89 %
		LR	88,96 %
		DT	91,57 %
Pasiūlytas metodas		LSVM^{PSO}_30K_SpeedUP	90,14 %
		LSVM$_3^{PSO}$_30K_SpeedUP	90,52 %
		LSVM^{PSO}	93,17 %

Rezultatai parodė, kad pakankamą tikslumą galima pasiekti naudojant mažesnę mokymosi duomenų aibę. Pasiūlytas metodas parodė geresnius rezultatus lyginant su Rain (2013) bei Shaikh ir Deshpande (2016), kai buvo naudojamas su didžiausia duomenų aibe Books bei KindleStore duomenų aibe. Pasiūlytas metodas taip pat parodė geresnius rezultatus lyginant su konvoliuciniais neuroniniais tinklais (angl. *Convolutional neural network*; CNN), kontekstinės faktorizacijos mašina (angl. *Contextual factorization machine*; CFM) ir poziciją žinančią faktorizacijos mašina (angl. *Position-aware factorization machine*; PFM), kai buvo naudojamas su Phones&Accessories duomenų aibe lyginant su Wang ir kt. (2018). Vis dėlto pasiūlytas metodas šiek tiek nusileido klasikiniam LSVM, kai buvo naudojamas su Electronics ir Phones&Accessories duomenų aibėmis, lyginant su Haque ir kt. (2018). Rezultatai parodė, kad pasiūlytas metodas gali būti naudojamas kaip alternatyva klasikiniams ir moderniems kitų autorių naudojamiems metodams, naudojant dideles tekstinių duomenų aibes.

Realių duomenų eksperimentų ciklas

3.3 lentelėje pateikiami klasikinio LSVM, LSVM$_{RS}$ (parametrų derinimui naudota atsitiktinė paieška), LSVM$_{Bopt}$ (parametrų derinimui naudota Bajeso optimizacija) ir pasiūlyto metodo LSVM^{PSO}_s_SpeedUP rezultatai. Šie rodo, kad tiksliausias yra

$LSVM_{RS}$ su Person duomenų aibe. Tikslumo skirtumas lyginant $LSVM_{RS}$ su pasiūlytu metodu nėra didelis – 0,1 %; lyginant su $LSVM_{Bopt}$ – 1,22 %.

3.3 lentelė: Realių duomenų eksperimentų ciklo rezultatai

Metodas	ACC	PPV	NPV	TPR	TNR	$F_1 score$	AUC
			Person				
LSVM	72,51 %	59,30 %	76,90 %	**45,76 %**	85,13 %	51,57 %	74,43 %
$LSVM_{RS}$	**76,20 %**	**81,00 %**	75,57 %	34,13 %	**96,05 %**	47,82 %	**76,21 %**
$LSVM_{Bopt}$	74,98 %	69,82 %	**77,05 %**	42,72 %	90,21 %	**51,81 %**	75,71 %
$LSVM^{PSO}\_287\_$	76,10 %	76,68 %	76,39 %	38,59 %	93,79 %	50,93 %	75,91 %
			Event				
LSVM	70,70 %	58,83 %	**76,04 %**	**51,54 %**	80,81 %	**54,70 %**	73,97 %
$LSVM_{RS}$	71,16 %	62,53 %	74,39 %	43,56 %	85,74 %	50,41 %	74,41 %
$LSVM_{Bopt}$	**71,89 %**	63,21 %	75,26 %	46,15 %	85,48 %	52,55 %	**74,55 %**
$LSVM^{PSO}\_300\_$	**71,89 %**	**63,78 %**	75,06 %	45,38 %	**85,89 %**	52,21 %	74,49 %

Pabraukimas „_“ reiškia, kad pabaigoje turi būti pridėta SpeedUP.

Vis dėlto $LSVM_{RS}$ nusileido pasiūlytam metodui ir $LSVM_{Bopt}$ su Event duomenų aibe, kurios duomenyse buvo daugiau dviprasmybių. Tikslumo skirtumas yra 0,73 % lyginant su pasiūlytu metodu ir $LSVM_{Bopt}$. Rezultatai parodė, kad $LSVM^{PSO}$_s_SpeedUP pralenkė klasikinį LSVM su abiem duomenų aibėmis (3,59 % su Person ir 1,19 % su Event).

4. BENDROSIOS IŠVADOS

1. Pasiūlytas hibridinis tekstinių duomenų metodas nuomonių analizei gali pagreitinti klasikinių mašininio mokymosi algoritmų, tokių kaip LSVM, LR, DT ir RF, klasifikavimą iki 4,7x–634,8x kartų, prarandant 0,29 %–4,06 % klasifikavimo tikslumo. Metodą sudaro:
    - **SpeedUP** – pagrindinė pasiūlyto hibridinio metodo dalis, kurios tikslas yra pagreitinti klasikinių mašininio mokymosi algoritmų klasifikavimą. Greitis padidėjo: LSVM atveju (iki 50,7x), LR atveju (iki 4,7x), DT

atveju (iki 634,8x) ir RF atveju (iki 72,3x). Tikslumo praradimai lyginant su klasikiniais ML algoritmais: LSVM (1,99 %–2,42 %), LR (1,91 %–1,97 %), DT (3,28 %–4,06 %) ir RF (2,35 %–3,00 %).

- **k-Means klasterizavimas** – atsakingas už mokymosi duomenų parinkimą. Tikslumo praradimai lyginant su klasikiniais ML algoritmais: LSVM (1,84 %–2,22 %), LR (1,82 %–1,92 %), DT (1,97 %–3,70 %) ir RF (1,84 %–1,92 %).
- **PSO derinimas** – atlieka hiperparametrų derinimą. Tikslumo praradimai lyginant su klasikiniu LSVM (1,13 %–1,40 %).
- **Ansamblis** – atlieka metodų kombinavimą ir balsavimą. Tikslumo praradimai lyginant su klasikiniais ML algoritmais: LSVM (0,29 %–0,59 %), LR (1,12 %–1,56 %), DT (tikslumas didesnis lyginant su klasikiniu DT 2,0 %–2,97 %) ir RF (tikslumas didesnis lyginant su klasikiniu RF 0,69 %–2,16 %).

2. Remiantis eksperimentų rezultatais, pasiūlyti du rekomenduojami parametrų rinkiniai hibridiniam metodui: pirmas rinkinys autonominiam metodui ir antras – ansambliui. Geriausi rezultatai pasiekti, kai LSVM yra naudojamas su pirmu (tikslumas padidėjo 0,96 %–1,02 %) bei antru (tikslumas padidėjo 1,70 %–1,83 %) rinkiniais. Rekomenduojama išjungti PSO derinimo metodą ansamblio atveju, kadangi pastarojo geriausi rezultatai pasiekiami naudojant silpnesnius metodus – pralenkė ansamblį su PSO derinimo metodu 0,02 %–0,26 %.
3. Remiantis rezultatais, gautais palyginus pasiūlytą hibridinį metodą su kitų autorių darbais, nustatyta, kad pasiūlytas metodas gali būti naudojamas kaip alternatyva klasikiniams (LSVM, LR, MNB, RF, DT) (tikslumas mažesnis 0,35 %–3,00 %, tikslumas didesnis 0,28 %–9,86 %) ir moderniems (CNN, CFM, PFM) (tikslumas didesnis 6,77 %–

17,82 %) kitų autorių naudojamiems metodams, naudojant dideles tekstinių duomenų aibes ir nenaudojant galingų kompiuterių. Be to, PSO parametrų derinimo metodas gali būti alternatyva tokiems populiariems metodams kaip atsitiktinė paieška (tikslumas mažesnis – 0,10 %, tikslumas didesnis – 0,73 %) ir Bajeso optimizacija (tikslumas didesnis – 1,12 %), klasifikavimui naudojant realius duomenis viešosios nuomonės tyrimuose.

PUBLIKACIJŲ SĄRAŠAS

Straipsniai recenzuojamuose periodiniuose moksliniuose žurnaluose:

1. Korovkinas, K., Danėnas, P., Garšva, G., 2017. SVM and Naïve Bayes Classification Ensemble Method for Sentiment Analysis. Baltic Journal of Modern Computing, 5(4), pp. 398–409.
2. Korovkinas, K., Danėnas, P., Garšva, G., 2019. SVM and k-Means Hybrid Method for Textual Data Sentiment Analysis. Baltic Journal of Modern Computing, 7(1), pp. 47–60.
3. Korovkinas, K., Danėnas, P., Garšva, G., 2020. Support Vector Machine Parameter Tuning Based on Particle Swarm Optimization Metaheuristic. Nonlinear Analysis: Modelling and Control, 25(2), pp. 266–281.

Straipsniai recenzuojamuose konferencijų leidiniuose:

1. Korovkinas, K., Danėnas, P., Garšva, G., 2018. SVM Accuracy and Training Speed Trade-Off in Sentiment Analysis Tasks. In International Conference on Information and Software Technologies, Springer, Cham, pp. 227–239.
2. Korovkinas, K., Garšva, G., 2018. Selection of Intelligent Algorithms for Sentiment Classification Method Creation. Proceedings of the International Conference on Information Technologies, Vol–2145, Kaunas, Lithuania, pp. 152–

157, ISSN 1613-0073, CEUR. Prieiga internete: http://ceur-ws.org/Vol-2145/p26.pdf

3. Vaitonis, M., Masteika, S., Korovkinas, K. 2018. Algorithmic Trading and Machine Learning Based on GPU. Proceedings of the Symposium for Young Scientists in Technology, Engineering and Mathematics, Vol–2147, Gliwice, Poland, pp. 49–54, ISSN 1613-0073, CEUR. Prieiga internete: http://ceur-ws.org/Vol-2147/p09.pdf
4. Korovkinas, K. 2020. A Hybrid Method for Textual Data Classification Based on Support Vector Machine with Particle Swarm Optimization Metaheuristic and k-Means Clustering. Proceedings of Baltic DB&IS 2020 Doctoral Consortium, Vol–2620, Tallinn, Estonia, pp. 81–88, ISSN 1613-0073, CEUR. Prieiga internete: http://ceur-ws.org/Vol-2620/paper11.pdf

Santraukos konferencijų leidiniuose:

1. Korovkinas, K., Garšva, G., 2018. Large Scale Sentiment Analysis Using NLP Based Feature Extraction Technique and PSOLinearSVM. Data analysis methods for software systems: 10th international workshop, Druskininkai, 2018. ISBN 978-609-07-0043-3. Prieiga internete: https://www.journals.vu.lt/proceedings/article/view/12634/11171

TRUMPOS ŽINIOS APIE AUTORIŲ

Konstantinas Korovkinas baigė informatikos studijų bakalauro programą Viniaus Gedimino technikos universitete (2007 m.). 2014 m. baigė verslo informatikos magistro studijas Vilniaus universitete ir įgijo informacijos sistemų magistro laipsnį (Magna Cum Laude). 2015–2019 metais studijavo Vilniaus universiteto Kauno fakulteto doktorantūroje (gamtos mokslai, informatika). Nuo 2018 metų Vilniaus universiteto Kauno fakultete dėsto kompiuterių architektūrą ir „Python“ programavimą.

INTRODUCTION

Research context

Textual data sentiment analysis (SA) became very popular when people started using the Internet, to be more concrete when e-shops and social networks, blogs and other platforms appeared where people could write their comments. Nowadays if you want to get an opinion about surveys, social, economic and others events, you can find all information you need on the Internet. The main goal of research related to sentiment analysis is to obtain authors' feelings expressed in positive or negative comments. This analysis is performed at multiple levels: document, sentence, and aspect. According to Pang and Lee, the term "sentiment" appeared in papers in 2001 and subsequently in papers in 2002. Opinion mining is another term in certain respects parallel to sentiment analysis that appeared in the 2003 paper by Dave et al. They described an ideal opinion-mining tool that "would process set of search results for the given item, generating a list of product attributes (quality, features, etc.) and aggregating opinions about each of them (poor, mixed, good)". Liu gives a definition of sentiment analysis. He described it as "the field of study that analyzes people's opinions, sentiments, appraisals, attitudes, and emotions toward entities and their attributes expressed in written text". According to him, this field represents a huge problem space due to "many related names and slightly different tasks". "Sentiment analysis, opinion mining, opinion analysis, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, and review mining", according to Liu, are "all under the umbrella of sentiment analysis". Basically, sentiment analysis is divided into lexicon-based methods and machine learning (ML) methods. The authors also mixed aforementioned methods to achieve better results. The lexicon-based approach involves calculating orientation for a document from the semantic

orientation of words or phrases in the document. At the present time the popularity of sentiment analysis has greatly increased and is still growing, because of a huge amount of text data, which is accessible on the Internet and could be very useful for companies, users etc. The usage of lexicon-based methods for a huge amount of data extracted from social media websites is not very effective, because of their unstructured format and the data can contain textual peculiarities, informal and dynamic nature of language, new slang, abbreviations, and new expressions. Also, it could significantly increase computational costs. However, this dissertation investigates sentiment classification using machine learning methods.

Research problem

Sentiment analysis of text is considered as a very challenging area: although a lot of work has been done in this field, accuracy (ACC) is still rather average due to comments, slang, smiles etc. We need to understand the whole context of the sentence, because even a single word can change the polarity of a sentence, and this might have a significant impact especially in currently sensitive domains such as medicine, business etc. Another problem which is faced in this area is a huge data amount. If a customer needs to get an opinion about a product, hotel, flight etc. – it becomes a very hard task. In the case of a large volume of data the performance of machine learning algorithms decreases depending on the dataset size – the higher the number of features is, the longer computation time it requires. In order to solve this problem, researchers use various techniques: parallelism, implementation on graphics processor unit, using cloud computing technology, selecting only representative data for training etc. This leads to another problem – the need for special hardware or a more powerful computer, cloud provider with required software etc. This dissertation investigates the problem of big computational costs achieved by machine learning algorithms on large scale textual

datasets. Therefore, the main focus is on increasing execution speed without or with a slight loss of accuracy and without the need of special software or hardware.

Object of research

The main object of this research is the textual data classification methods, their execution speed and accuracy in large scale datasets, sentiment analysis.

Goal and objectives of the research

The goal of the research is to propose a hybrid sentiment analysis method with a recommended set of parameters for large textual data with a better execution time and with a similar classification accuracy compared with classical methods.

The objectives of the dissertation are:

1. To propose a hybrid method that increases the classification speed of the selected classical machine learning algorithms, which are commonly used for textual data sentiment analysis with a similar classification accuracy.
2. To perform the experimental evaluation of the method proposed and select the recommended set of parameters for it, as well as to improve classification accuracy.
3. To perform a comparison of the proposed hybrid method with other authors' works on large scale textual datasets and evaluate obtained results.

Research methodology and tools

The following methods were used:

1. Bibliographic research on the sentiment analysis field helped to formulate the research tasks and goals.
2. Analysis of related works helped to select machine learning algorithms for the proposed hybrid method.
3. Proposed hybrid method for textual data sentiment analysis is described in Chapter "Methodology of the research".

4. Experimental research methodology and experimental research for comparative analysis of the proposed method are described in Chapter "Experiments and results".
5. Formulation of conclusions after each chapter and general conclusions at the end of research.

For the proposed hybrid method development and performing experimental research were used: Python programming language, scikit-learn[7] – library for machine learning. For the dissertation LaTeX[8], a document preparation system was used; for the presentation of graphical results and diagrams latex Tikz[9] package was employed.

Scientific novelty

In this dissertation a hybrid method for textual data sentiment analysis with a recommended set of parameters suitable for large scale datasets is proposed. The method consists of the following:

1. SpeedUP method – this is the main part of the proposed hybrid method, whose aim is to increase the classification speed of classical machine learning algorithms.
2. k-Means clustering method – this method is responsible for training data selection.
3. PSO tuning method – this method performs hyperparameters tuning for linear support vector machine (LSVM).
4. Ensemble method – this is the last part of the proposed hybrid method, which performs combination and voting of the machine learning algorithms.

In this dissertation, machine learning algorithms for textual data sentiment analysis were reviewed and five most common

[7] `https://scikit-learn.org/`

[8] `https://www.latex-project.org/`

[9] `https://www.overleaf.com/learn/latex/TikZ_package`

were selected. Effectiveness metrics, average ranking and statistical significance were performed with selected machine learning algorithms.

Based on experimental results the recommended set of parameters for the proposed hybrid method was selected and presented. The results showed that the proposed method increased the classification speed of classical machine learning algorithms on large scale textual datasets with slight loses in terms of accuracy; it is also competitive with state-of-the-art methods.

Unlike the techniques proposed previously, SpeedUP automatically performs all parts of the proposed hybrid method, based on specified parameters, which are recommended in this research and are set as default in the SpeedUP method. Depending on determined subset size, the size of training data is calculated; depending on it the k-Means method automatically selects the appropriate amount of training data (in the case of an ensemble method it selects as many training datasets as the selected number of classifiers) and passes it to LSVM input; if the k-Means method is switched off datasets will be selected randomly. PSO automatically selects a C parameter for LSVM and its training is performed with this parameter; in the case of ensemble the same number of C parameters and classifiers is also selected. All calculations, training data selection, testing dataset dividing into subsets, tuning of the hyperparameters, joining ML algorithms to ensembles, joining of results and voting are performed automatically by the SpeedUP method.

The proposed method can be applied to classify textual data sentiments and is suitable to work with large scale datasets without using supercomputers.

Practical significance

Since sentiment analysis of text is still a very challenging area and at the same time it is very widely applicable in practice for product reviews, customer churn prediction, fraud detection,

election etc., the proposed hybrid method could be:

1. Successfully applied in these areas for the development of new models or for improving the existing ones.
2. Be suitable for work with large scale textual datasets and allows classifying sentiment without using high-performance computers.
3. Be useful with surveys, because opinions are taken from social networks, articles comments etc., which allows forming different opinions on the current topic.

Defended statements

1. The proposed hybrid method for textual data sentiment analysis in large scale datasets can increase the classification speed of classical machine learning algorithms such as linear support vector machine, logistic regression (LR), decision tree (DT) and random forest (RF), whereas losses in terms of accuracy are not very great.
2. The best results are achieved when the proposed method with the recommended set of parameters is used with LSVM, compared with those when it is employed with multinomial naïve Bayes (MNB), logistic regression, decision tree and random forest.
3. The proposed hybrid method could be used as an alternative method on large scale textual datasets for classical and state-of-the-art methods which are used by other authors. Moreover, PSO tuning method could be competitive with such popular methods as random search (RS) and Bayesian optimization (Bopt).

Presentation and approbation of the results

The results of this research were presented at three international and two national scientific conferences. Three papers have been published in reviewed journals (one journal is included in ISI Web

of Science database), four in periodical conference proceedings and one abstract in conference abstracts proceedings.

Structure of the dissertation

The dissertation consists of four chapters, general conclusions, a list of references and appendices. The scope of dissertation is 170 pages including 55 tables and 33 figures. The list of references contains 212 various sources, including books, scientific papers, patents, technical reports and Internet sources.

GENERAL CONCLUSION

1. The proposed hybrid method for textual data sentiment analysis in large scale datasets can increase classification speed up to 4.7x-634.8x of classical machine learning algorithms such as LSVM, LR, DT and RF, while losing in terms of accuracy is 0.29%-4.06%. The method includes the following:
    - **SpeedUP** – the main part of the proposed hybrid method, whose aim is to increase the classification speed of classical machine learning algorithms. The speed increased: LSVM (up to 50.7x), LR (up to 4.7x), DT (up to 634.8x) and RF (up to 72.3x). Accuracy loses compared with classical ML algorithms were: LSVM (1.99%-2.42%), LR (1.91%-1.97%), DT (3.28%-4.06%) and RF (2.35%-3.00%).
    - **k-Means clustering** – responsible for training data selection. Accuracy loses compared with classical ML algorithms were: LSVM (1.84%-2.22%), LR (1.82%-1.92%), DT (1.97%-3.70%) and RF (1.84%-1.92%).
    - **PSO tuning** – performs the tuning of hyperparameters. Accuracy loses compared with classical LSVM were 1.13%-1.40%.
    - **Ensemble** – performs combination and voting of the

machine learning algorithms. Accuracy loses compared with classical ML algorithms were: LSVM (0.29%-0.59%), LR (1.12%-1.56%), DT (outperformed classical DT by 2.0%-2.97%) and RF (outperformed classical RF by 0.69%-2.16%).

2. Two recommended sets of parameters for the hybrid method were proposed according to the experimental results: the first for the standalone classifier and the second for ensemble. The best results were achieved, when LSVM was used with the first set (accuracy increased by 0.96%-1.02%) and the second set (accuracy increased by 1.70%-1.83%) of parameters. It is recommended to turn off PSO tuning if the ensemble method is enabled, since ensemble works better with weaker classifiers – it outperformed ensemble with PSO tuning by 0.02%-0.26%.
3. According to the results achieved during the comparison with other authors' work, the proposed hybrid method could be used as an alternative method on large scale textual datasets for classical (LSVM, LR, MNB, RF, DT) (the proposed method lost by 0.35%-3.00% and outperformed it by 0.28%-9.86%) and state-of-the-art methods (CNN, CFM, PFM) (the proposed method outperformed by 6.77%-17.82%), which are used by other authors on not very powerful computers. Moreover, PSO tuning applied on LSVM could be an alternative to such popular methods as random search (the proposed method lost by 0.10% and outperformed by 0.73%) and Bayesian optimization (the proposed method outperformed by 1.12%) on the real-world data classification tasks in public opinion research.

Konstantinas Korovkinas

HIBRIDINIS TEKSTINIŲ DUOMENŲ METODAS NUOMONIŲ ANALIZEI

Daktaro disertacijos santrauka
Gamtos mokslai
Informatika (N 009)
Redaktorė Jorūnė Rimeisytė – Nekrašienė

Konstantinas Korovkinas

HYBRID METHOD FOR TEXTUAL DATA SENTIMENT ANALYSIS

Summary of a Doctoral Dissertation
Natural Sciences
Informatics (N 009)
Editor Zuzana Šiušaitė

Vilniaus universiteto leidykla
Saulėtekio al. 9, III rūmai, LT-10222 Vilnius
El. p.: info@leidykla.vu.lt, www.leidykla.vu.lt
Tiražas 30 egz.