ISSN 1648-8776
JAUNŲJŲ MOKSLININKŲ DARBAI. Nr. 49 (2). 2019 DOI: 10.21277/jmd.v49i2.235

DIDELĮ KATEGORIJŲ KIEKĮ TURINČIŲ DRAUDIMO BENDROVĖS KLIENTŲ UŽKLAUSŲ, GAUTŲ ELEKTRONINIAIS LAIŠKAIS, LIETUVIŠKO TEKSTO KLASIFIKAVIMAS

Karolis Kiaunė, Simona Ramanauskaitė
Vilniaus Gedimino technikos universitetas
E. p.: karolis.kiaune@stud.vgtu.lt, simona.ramanauskaite@vgtu.lt

Įvadas

Šiandieniniame pasaulyje dauguma įmonių vis daugiau naudoja naujas technologijas. Šios technologijos dažnai padeda paspartinti įmonių procesus, automatizuoti rankų darbą. Viena iš didžiausių draudimo kompanijų Lietuvoje AB „Lietuvos draudimas“ nėra išimtis. Ši kompanija užima nemažą rinkos dalį, joje dirba daugiau nei 1100 darbuotojų, taip pat ši įmonė turi nemažai draudimo konsultantų, brokerių bei tarpininkų. Darbuotojams iškyla klausimų, pastebima klaidų su pagrindine sistema, prireikia papildyti informaciją sudarant sutartis, gaunama įvairių užklausų ir kt. Už visas šias užklausas įmonėje atsakingas skambučių centras. Paprastai šias užklausas skyriaus darbuotojai turi klasifikuoti patys, nuspręsdami, kokiai kategorijai užklausa priklauso pagal jos turinį. Bėgant metams ir įmonei plečiantis šių užklausų tik daugės. O tai reikalauja gana nemažai žmogiškųjų išteklių ir laiko.

Tyrimo tikslas – įvertinti didelį kategorijų kiekį turinčių draudimo bendrovės klientų užklausų, parašytų lietuvių kalba, klasifikavimo tikslumą, taikant skirtingus teksto apdorojimo ir klasifikavimo metodus.

Siekiant šio tikslo, yra keliami tokie ***tyrimo uždaviniai***:

1. Apžvelgti egzistuojančius teksto klasifikavimo metodus.
2. Išanalizuoti rastų klasifikavimo metodų tikslumą lietuviškam tekstui klasifikuoti.
3. Pasiūlyti tinkamiausią metodų derinį skambučių centro užklausoms automatiniu būdu klasifikuoti.

Rašytinės natūralios kalbos apdorojimo principai

Kalba – tai žodyno ir gramatikos sistema, padedanti žmogui atskleisti save ir jį supantį pasaulį sau ir kitiems. Tačiau tobulėjant technologijoms, atsirandant išmaniesiems telefonams, laikrodžiams, kurie gali turėti integruotus garsiakalbius, atsirado galimybė palengvinti žmogaus kasdienius darbus. Taip atsirado poreikis išmokyti kompiuterį suprasti tą pačią kalbą, kurią vartoja žmonės. Tačiau kompiuteris gali atlikti tik tuos veiksmus, kurie yra tikslingai jam suprogramuoti, todėl tai padaryti buvo gana sunku. Šį iššūkį padėjo spręsti dirbtinis intelektas. Dirbtinis intelektas – tai informatikos mokslo šaka, tirianti kompiuterių naudojimą panaudojant žmogaus proto savybes (analizavimas, išvadų darymas, galimybė mokytis iš patirties, kalbos atpažinimas ir kt.) (Valstybinė lietuvių kalbos komisija, 2005).

Dirbtinio intelekto taikymas natūralios kalbos apdorojimui

Viena iš dirbtinio intelekto panaudojimo sričių yra NLP (angl. *Natural language processing*) – natūralios kalbos apdorojimas. Šefildo universiteto natūralios kalbos apdorojimo tyrėjų grupė pateikia tokią NLP apibrėžtį: „NLP yra moderni šiuolaikinė kompiuterinė technologija ir kartu tai yra taikytinas metodas tyrimams ir vertinimams apie pačią žmonių kalbą atlikti“ (Monkus, 2010). Kalba gali būti 2 tipų, t. y. rašytinė ir skaitytinė. Turint kompiuteriu parašytą tekstą, dažniausiai sprendžiamos tokios pagrindinės NLP problemos (Manning, Raghavan, & Schütze, 2009):

1. Teksto klasifikavimas (angl. *Text Classification*) – užduotis priskirti iš anksto nustatytas kategorijas tekstiniams dokumentams. Teksto klasifikavimas gali pateikti konceptualius dokumentų rinkinių vaizdus, galimus pritaikyti realiame pasaulyje. Pavyzdžiui, naujienų istorijos yra paprastai skirstomos pagal temas, akademiniai straipsniai dažnai klasifikuojami pagal techninius duomenis. Taip pat kitas teksto klasifikavimas yra taikomas laiškuose, kai norima atskirti brukalą (angl. *spam*) ir naudingo turinio laiškus (Yang & Joachims, 2008).
2. Mašininis vertimas (angl. *Machine Translation*) – užduotis, kai įvestis susideda iš simbolių sekos tam tikra kalba, o kompiuterio programa turi

paversti ją simbolių seka kita kalba. Tai paprastai taikoma natūralioms kalboms, pvz., versti iš anglų į prancūzų kalbą (Goodfellow, Bengio, & Courville, 2016).

3. Dokumentų apibendrinimas (angl. *Document Summarization*) – užduotis sukurti įvesties teksto santrauką, kurioje būtų aprašoma pagrindinė originalaus teksto prasmė. Geriausios tokio tipo sistemos naudoja ištraukimo metodus, kurie iškerpa ir sujungia teksto dalis, taip sukurdamos suglausto teksto versiją (Rush, Chopra, & Weston, 2015).
4. Klausimų atsakymas (angl. *Question Answering*) – sistema, kuri sugeba apdoroti įvesties informaciją ir prisiminti svarbiausius faktus, sukurtus taip, kad būtų galima juos vėliau pasiimti ir jais argumentuoti. Toks sistemos gebėjimas išlieka didelė problema, kurią galima išspręsti tik konkrečioje aplinkoje. Šiuo metu geriausias būdas atsiminti ir išgauti specifinius konstruojamuosius faktus yra tikslinės atminties (angl. *explicit memory*) mechanizmo naudojimas (Goodfellow, Bengio, & Courville, 2016).

Terminų svorio parinkimo metodai teksto klasifikavimui

Prieš taikant algoritmus, tekstui reikalinga transformacija, kuri paverstų jį į skaitinių požymių rinkinį, kuris būtų priimtinas algoritmams. Paprastai tekstas pateikiamas kaip tam tikrų teksto elementų svarbą atspindinčių svorių vektorius. Dažniausiai naudojami savybių matricos parinkimo būdai yra žodžių maišas ir fiksuoto ilgio vektorius (Balys, 2009).

Paprasčiausias metodas yra priskirti svorį, kuris lygus termino *t* pasirodymo skaičiui dokumente *d*. Tokia svorių priskyrimo schema vadinama terminų dažnumu ir žymima $tf_{t,d}$ (Jurafsky & Martin, 2008; Manning et al., 2009).

Dokumento *d* svoriai, aprašomi pagal *tf* svorius, gali būti vertinami kaip šio dokumento kiekybinis apibendrinimas. Toks dokumento aprašymas vadinamas žodžių maišo modeliu, kur tikslus terminų rikiavimas neturi įtakos, bet termino pasirodymų skaičius yra svarbus. Tokiu atveju paimama informacija tik apie žodžių pasirodymo skaičių. Taigi dokumentai „Mary is quicker than John“ ir „John is quicker than Mary“ bus vienodi (Manning et al., 2009).

Lietuviško teksto klasifikavimo tyrimai

Nors natūralios kalbos teksto klasifikavimo sprendimai jau pakankamai išvystyti anglų kalbai, lietuvių kalbai jie nėra tinkami dėl pastarosios savitumo ir sudėtingumo. Todėl natūralios kalbos tekstų klasifikavimas lietuvių kalba rašytiems tekstams vis dar yra aktuali problema ir reikalauja papildomų tyrimų.

Vytauto Mickevičiaus ir kt. straipsnyje „Classification of Short Legal Lithunian Texts“ aprašomas tyrimas, kurio tikslas – ištirti Seimo narių pasisakymus parlamente. Šio tyrimo metu buvo eksperimentuojama su skirtingų savybių matricų metodais, pvz., žodžių maišas (angl. *bag-of-words*), *tf-idf*, *n-gramų* (angl. *n-grams*) skirtingais klasifikatoriais: atraminių vektorių mašinos (SVM), naudojant skirtingus branduolius ir multinominę logistinę regresiją. Šiame darbe siekiant išvengti funkcinių žodžių ir simbolių įtakos, buvo atlikti normalizavimo darbai (Mickevičius ir kt., 2015):

1. panaikinti punktuacijos ženklai ir skaičiai;
2. didžiosios raidės paverstos mažosiomis;
3. 185 nereikšmingi (pvz.: *ir*, *ar* ir t. t.) iš 3299 unikalių žodžių buvo panaikinti.

Tyrimo metu buvo naudojamos tik 7 kategorijos ir nustatyta, kad bag-of-words metodu geriausias klasifikavimo tikslumas buvo 72 %, 3-gram metodu – 68 %, o tf-idf metodu – 83 %. Tai geri rezultatai, tačiau reikėtų pažymėti, kad analizuojami tekstai yra pakankamai formalūs ir specifiniai, todėl skirtumai tarp klasių gana aiškūs.

Kitame straipsnyje „Automatic Thematic Classification of the Titles of the Seimas Votes“ V. Mickevičius ir kt. tiria panašią problemą, kaip ir prieš tai nagrinėtame straipsnyje – Seimo narių balsavimų duomenis. Pritaikytos panašios duomenų normavimo ir gavimo taisyklės ir pasitelktos tos pačios 7 klasės, savybių matricos šiam tyrimui buvo žodžių maišas ir *n*-gramos, o teksto klasifikatoriai buvo parinkti SVM ir K artimiausių kaimynų (angl. *K Nearest Neighbors*, toliau – k-NN) metodai (Mickevičius ir kt., 2015).

Tyrimo metodologija

Tyrimui naudojami duomenys

Tyrimui pasirinkta Lietuvoje veikianti draudimo bendrovė, kuri per metus aptarnauja apie pusę milijono klientų. Dauguma jų yra privatūs klientai, kuriems dažnai kyla kokių nors klausimų. Tokių užklausų apdorojimą dažniausiai atlieka Skambučių centro skyrius. Šiame skyriuje užklausos apdorojamos tiek laiškais, tiek skambučiais. Kadangi šio tyrimo metu analizuojamas teksto klasifikavimas, tai tyrimui bus naudojamos tik Skambučių centro iš klientų gautos užklausos elektroniniais laiškais.

Per 2018 metus šis skyrius sulaukė beveik 90 tūkstančių laiškų, kurie buvo pasiskirstę per įvairias pašto dėžutes. Kiekviena pašto dėžutė gali turėti tam tikrus katalogus. Katalogai gali identifikuoti, su kokia platesne tema susiję elektroniniai laiškai, arba nustatyti laiškų svarbumą. Toliau kiekvienas laiškas gauna tam tikrą kategoriją, kuri atspindi, su kokia tema susijęs laiškas ar kokių veiksmų reikės imtis norint jį apdoroti toliau.

Yra naudojamos 5 pašto dėžutės, į kurias per dieną ateina apie 500 elektroninių laiškų (kiekvieno-

je dėžutėje nuo 20 iki 250 laiškų). Dažnai pasitaiko atvejų, kai žmonės susimaišo ir išsiunčia ne į tą dėžutę užklausas, todėl yra apibrėžtos 33 kategorijos visose 5 dėžutėse. Šios kategorijos nusako visus užklausų galimus atvejus. Šios kategorijos priskiriamos žmogaus, kuris apdoroja gautą užklausą, ir toliau bus naudojamos šiame tyrime. Esamų elektroninių laiškų rankinio sužymėjimo tikslumo užtikrinimui remtasi ne pačiais naujausiais, o mėnesio ir senesniais elektroniniais laiškais. Jei jie netyčia būtų priskirti ne tai kategorijai, darbo metu netinkamas priskyrimas iškiltų ir būtų pakeistas, priskiriant tą elektroninį laišką atitinkamai kategorijai. Nors visiško tikslumo užtikrinti neįmanoma, vadovaujamasi prielaida, kad jei įmonė taip klasifikuoja elektroninius laiškus, tai tokios pačios sistemos turėtų laikytis ir automatinis elektroninių laiškų klasifikatorius.

Tyrimo eiga

Tyrimo metu nuskaitomi elektroninių laiškų tekstai iš visų penkių pašto dėžučių. Gavus neapdoroto teksto rezultatus, toliau tyrimas vykdomas atliekant 7 bandymus:

1. Panaikinami įmonės saugumo pranešimai ir laiškų pasirašymai. Tai laiško apačioje esantis papildomas tekstas, kuris nesusijęs su pagrindiniu laiško turiniu ir skirtas rašančiojo kontaktams ar informacijai apie negalimą informacijos viešinimą nurodyti. Jis pašalinamas remiantis tarpu nuo pagrindinio teksto ar šio teksto bloko išskirtiniu formatavimu.
2. Pašalinami ne tokie reikšmingi kategorizavimo požiūriu lietuviški žodžiai, pavyzdžiui, prielinksniai, dalelytės, jungtukai, jaustukai, ištiktukai.
3. Sutrumpinami žodžiai, paliekant tik jų standartines formas. Tam naudojamas Kompiuterinės lingvistikos centro sukurtas morfologinis anotatorius (*Morfologinis anotatorius internete*), kuris pateiktą tekstą išanalizuoja ir pateikia kiekvieno žodžio (ar jų junginio) lemą, standartinę formą.
4. 1 ir 2 bandymo kombinacija. Šios kombinacijos metu analizuojamas tik elektroninio laiško tekstas (be kontaktinės ir kitos nesusijusios informacijos), kuriame pašalinti ne tokie reikšmingi lietuviški žodžiai.
5. 1 ir 3 bandymo kombinacija. Šioje kombinacijoje ne tokie reikšmingi lietuviški žodžiai paliekami, bet visas tekstas keičiamas, žodžius pateikiant jų standartine forma.
6. 2 ir 3 bandymo kombinacija. Šiuo atveju elektroninio laiško turinys paliekamas pilnas, nešalinant autoriaus kontaktų ar saugos pranešimų laiško apačioje. Laiško tekste pašalinami nereikšminiai lietuviški žodžiai, o likęs tekstas pakeičiamas, kiekvieną žodį ar jų junginį pateikiant standartine forma.
7. 1, 2, 3 bandymo kombinacija. Sujungiami visi trys metodai, kada pašalinami autoriaus kontaktai laiško apačioje, eliminuojami nereikšminiai lietuviški žodžiai, likę žodžiai pakeičiami standartinėmis formomis.

Dėl įmonės vidaus politikos visų kategorijų pavadinimai negali būti atskleisti, bet jie apima praktiškai visą draudimo bendrovės darbinių elektroninių laiškų sritį: kreipimaisi dėl skirtingo tipo žalų, užklausos dėl informacijos pateikimo, atnaujinimo ir pan. Kadangi kategorijų kiekis yra gana didelis, t. y. 33, tai visi bandymai taip pat atliekami su pamažintu iki 19 kategorijų kiekiu. Šis kategorijų mažinimas atliekamas sujungiant kategorijas, kurias įmanoma, į platesnes. Analizuojant bandymo rezultatus pasirinkta naudoti kryžminio tikrinimo metodą, kai k yra 3, 5, 10. Lankstymų skaičius k nusako, į kiek dalių bus dalijami duomenys ir paskirstomi algoritmo apmokymui bei testavimui. Lankstymų skaičiaus reikšmės k pasirinktos tokios, nes k=3 buvo siūloma senesnėje algoritmo įgyvendinimo versijoje kaip rekomendacija, k=5 siūlomas naujesnėje versijoje, o k=10 siūlomas teoriniuose šaltiniuose. Todėl su visomis rekomenduojamomis reikšmėmis ir atliekami testai.

Remiantis panašių tyrimų pavyzdžiais pastebėta, kad dauguma terminų svorių pasirinkimui naudoja *tf-idf* metodą. Lyginant *tf-idf* metodą su žodžių maišo metodu, matomas didesnis pranašumas naudojant *tf-idf* metodą, nes šis metodas nesusiduria su viena iš pagrindinių problemų, kada visi terminai laikomi vienodai svarbiais, vertinant užklausos tinkamumą. Todėl šiame tyrime buvo naudojamas terminų svorių parinkimas, naudojant *tf-idf* metodą.

1 pav. Tyrimo eigos schema

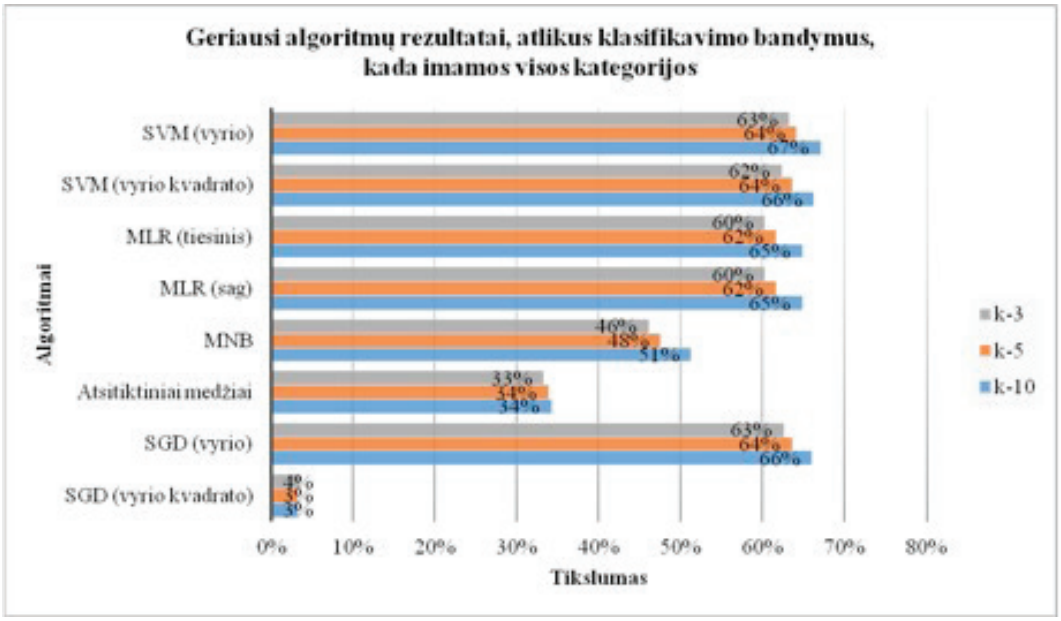
Užklausų automatiniam klasifikavimui buvo pasirinkti 5 prižiūrimo mokymosi algoritmai:

1. Naivaus bajeso (angl. *Naive Bayes*) algoritmas – nagrinėtuose šaltiniuose dažnai pasitaikantis siūlymas naudoti panašioms problemoms spręsti ir parodantis gana gerus rezultatus.
2. Atraminių vektorių mašinos (angl. *Support Vector Machines*) algoritmas. Šis algoritmas taip pat gana dažnas pasiūlymas literatūroje, kaip gerus rezultatus parodantis sprendimas. Branduolio funkcija buvo parinkta nemodifikuota pagrindinė linijinė branduolio funkcija. Taip pat naudojamos kelios nuostolių funkcijų modifikacijos, t. y. vyrio (angl. *hinge*) ir vyrio kvadrato (angl. *squared-hinge*). Šis algoritmas parodė geriausius rezultatus nagrinėjant Seimo narių balsavimus.
3. Logistinė regresija (angl. *Logistic Reegression*). Šis algoritmas nedaug kuo skiriasi nuo atraminių vektorių mašinų algoritmo. Kadangi bus nagrinėjama daugiau nei viena savybė, šiame straipsnyje buvo pasirinkta naudoti multinominė logistinė regresija. Taip pat šiam algoritmui buvo taikomos kelios skirtingos optimizavimo problemą sprendžiančios funkcijos, t. y. *sag* (pagal dokumentaciją, ši funkcija labiau tinka didesniems duomenų kiekiams) ir *liblinear* (kūrėjų numatytas metodas).
4. Stochastinio gradiento nusileidimo (angl. *Stochastic Gradient Descent*) algoritmas. Nors literatūroje jis yra nedažnai naudojamas ir nėra toks populiarus, tačiau jame galima pasirinkti nemažai skirtingų nuostolių funkcijų, kurios remiasi tokiais algoritmais kaip logistinės regresijos arba atraminių vektorių mašinų algoritmu, todėl galima tikėtis gerų rezultatų. Taip pat bus naudojamos šio metodo kelios nuostolių funkcijų modifikacijos, t. y. vyrio ir vyrio kvadrato.
5. Atsitiktinio miško (angl. *Random Forest*) algoritmas. Nors literatūroje šis algoritmas dažnai neša labai prastus rezultatus, tačiau nėra aišku, kaip jis suveiks su laiškų klasifikavimu ir lietuvių kalba.

Šiame tyrime bus analizuojama, kaip šie teksto paruošimo ir klasifikavimo metodai veikia su lietuvių kalba rašytu tekstu, kuris turi pakankamai didelį kategorijų skaičių. Visa elektroninių laiškų klasifikavimo sistema realizuota *Python* programavimo kalba ir jos biblioteka *scikit-learn*, kurioje realizuoti pasirinkti ir kiti papildomi dirbtinio intelekto metodai. Kadangi ši biblioteka plačiai naudojama, galima pasikliauti jos algoritmų realizavimo tinkamumu ir efektyvumo užtikrinimu.

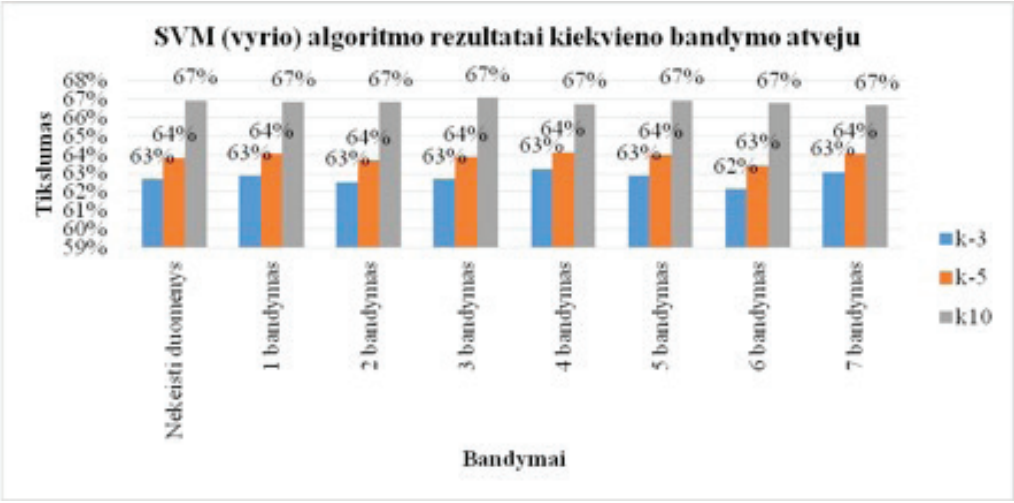
Tyrimo rezultatai

Analizuojant gautus tyrimo rezultatus pastebėta, kad didžiausio tikslumo klasifikuojant lietuvišką tekstą galima pasiekti VSM (vyrio) metodu. Taikant šį klasifikavimo metodą į 33 kategorijas elektroninius laiškus galima suskirstyti iki 67 % tikslumu, o kada kategorijų skaičius sumažinamas iki 19, tikslumas siekia 71 % (žr. 2 pav.). Tačiau vertėtų pažymėti, kad daugelio algoritmų rodomi rezultatai yra gana artimi, labiau skiriasi tik naudojant MNB, atsitiktinių medžių ir SGD (vyrio kvadrato) algoritmus gauti rezultatai.



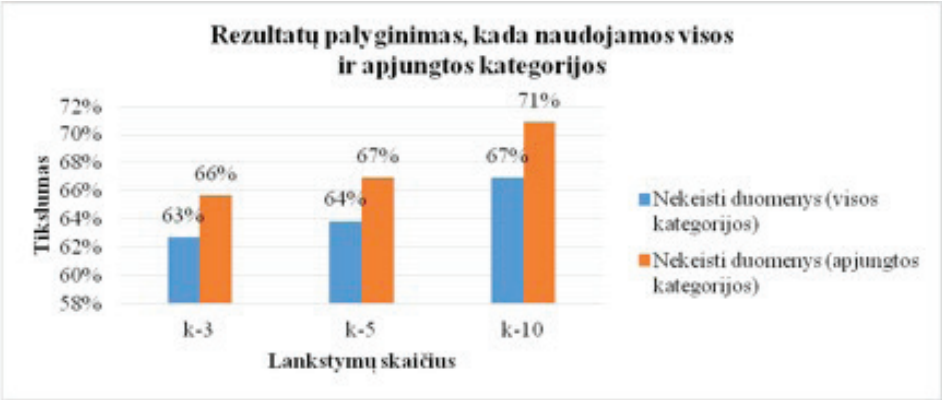
2 pav. Analizuotų teksto klasifikavimo algoritmų tikslumas tirtiems duomenims

Remiantis tyrimo rezultatais galima teigti, kad, atliekant pasirinktus duomenų valymo metodus (saugumo pranešimų šalinimą, laiškų parašų naikinimą, ne tokių reikšmingų klasifikavimui lietuviškų žodžių trynimą ar žodžių šaknų radimą), negaunamas didelis rezultatų gerėjimas lyginant su neapdorotu tekstu. Lyginant visus algoritmus tarpusavyje tiksliausiai visuose bandymuose pasirodė SVM algoritmas (2 pav.), todėl lyginant duomenų valymo metodus 3 paveiksle pateikiami tik naudojant šį algoritmą gauti rezultatai. Tačiau analogiška tendencija matoma ir kituose klasifikavimo metodų tyrimuose.



3 pav. Teksto paruošimo klasifikavimui bandymų tikslumo rezultatai SVM (vyrio) metodui

Reikėtų paminėti, kad apmokymui naudojant didesnę dalį duomenų gaunamas didesnis klasifikavimo tikslumas (naudojant 90 % duomenų apmokymui ir 10 % testavimui gaunamas iki 71 % tikslumas; naudojant 80 % duomenų apmokymui ir 20 % testavimui – iki 67 % tikslumas, o naudojant 67 % duomenų apmokymui ir 33 % testavimui – iki 66 % tikslumas). Taip pat pastebėta, kad, mažinant kategorijų skaičių, klasifikavimo tikslumas didėja (žr. 4 pav.). Tačiau kategorijų mažinimas yra nepageidaujamas šioje draudimo įmonėje, todėl tiksli priklausomybė tarp kategorijų kiekio ir tikslumo nebuvo analizuojama.



4 pav. Kategorijų kiekio įtaka klasifikavimo tikslumui

Kadangi analizuoti teksto klasifikavimo algoritmai parodė gana panašius tikslumo rezultatus, papildomai buvo įvertintas ir naudojamų algoritmų našumas (žr. 5 pav.).

5 pav. Tyrimo rezultatų palyginimas pagal jų tikslumo ir našumo reikšmes

Naudojamų algoritmų našumas buvo vertinamas programiniame kode įterpiant algoritmo pradžios ir pabaigos laikų fiksavimą. Kadangi visi algoritmai buvo naudojami iš *Python* bibliotekos *scikit-learn*, tai tikima, kad jie realizuoti vienodai kokybiškai ir rezultatai atitinka paties algoritmo, o ne jo realizacijos santykinį našumą.

Remiantis algoritmų našumo ir tikslumo santykiu, realaus laiko sistemose tinkamiausias yra SGD (vyrio) metodas, nes jis nuo SVM (vyrio) metodo tikslumu skiriasi nedaug, o yra greičiausias iš bent 60 % tikslumą pasiekiančių metodų.

Išvados

1. Išanalizavus egzistuojančius teksto klasifikavimo metodus nustatyta, kad lietuvių kalba pareng-tiems tekstams klasifikuoti autoriai dažniausiai taiko logistinės regresijos, atraminių vektorių mašinų, k-artimiausių kaimynų metodus. Tačiau literatūroje siūlomi ir naivaus bajeso, stochastinio gradiento nusileidimo bei pasirinkimo medžių algoritmai, kuriuos taikant pasiekiama neblogų rezultatų.
2. Atlikus tyrimą nustatyta, kad, naudojant atraminių vektorių mašinų, stochastinio gradiento nusileidimo ir multinominės logistinės regresijos metodus, gaunami tiksliausi rezultatai, klasifikuojant lietuvišką tekstą. Šie rezultatai buvo gauti panaudojus beveik 51 tūkstantį skambučių centro užklausų, kurios buvo suskirstytos į 33 kategorijas. Tyrimas buvo atliekamas pritaikant skirtingas teksto valymo kombinacijas naudojant visas 33 kategorijas ir sujungtas kategorijas (19 kategori-

jų). Kategorijos, kurių turinys yra glaudžiai susijęs, buvo sujungiamos į platesnes temas.

3. Atlikus tyrimą ir išnagrinėjus jo rezultatus paaiškėjo, kad tinkamiausias yra atraminių vektorių mašinų algoritmas, kai yra parinkta linijinė branduolio funkcija ir vyrio nuostolių funkcija. Algoritmas geriausiai veikia pritaikius tam tikrą duomenų apdorojimą, t. y. panaikinus įmonės saugumo pranešimus bei laiškų pasirašymus ir pašalinus ne tokius reikšmingus klasifikavimui lietuviškus žodžius (prielinksnius, dalelytes, jungtukus, ištiktukus, jaustukus). Pritaikius šią teksto apdorojimo kombinaciją, gaunamas nuo 63,26 iki 66,75 % tikslumas, priklausomai nuo k reikšmės.

Literatūra

1. Balys V., 2009, *Mokslinės terminijos matematiniai modeliai ir jų taikymas leidinių klasifikavime*, Vilnius: Technika.
2. Goodfellow I., Bengio Y. & Courville A., 2016, *Deep Learning.* s.l.: MIT Press.
3. Jurafsky D. & Martin J. H., 2008, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Prentice Hall.
4. Kompiuterinės lingvistikos centras. Morfologinis anotatorius internete, http://donelaitis.vdu.lt/main.php?id=4&nr=7_2
5. Kotsiantis S. B., 2007, Supervised Machine Learning: A Review of Classification Techniques. *Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*. 16 Liepa. P. 3–24.
6. Manning C. D., Raghavan P. & Schütze H., 2009, *Introduction to Information Retrieval.* Cambridge: Cambridge University Press.
7. Mickevicius V., Krilavicius T. & Morkevicius V., 2015, Classification of Short Legal Lithuanian Texts. *The 5th Workshop on Balto-Slavic Natural Language Processing*, Rugsėjis. P. 106–111.
8. Mickevičius V., Krilavičius T., Morkevičius V. & Mackutė-Varoneckienė A., 2015, Automatic Thematic Classification of the Titles of the Seimas Votes. *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, Gegužė. P. 225–231.
9. Rush A. M., Chopra S. & Weston J., 2015, A Neural Attention Model for Abstractive Sentence Summarization. 3 Rugsėjis.
10. Scikit-learn, 2007-2018, *Cross-validation: evaluating estimator performance,* https://scikit-learn.org/stable/modules/cross_validation.html
11. Xu K. et al., 2016, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. 16 Balandis.
12. Yang Y. & Joachims T., 2008. *Text categorization*, http://www.scholarpedia.org/article/Text_categorization

Summary

CLASSIFICATION OF THE LITHUANIAN TEXT OF EMAIL ENQUIERIES OF AN INSURANCE COMPANY WITH A BIG NUMBER OF CUSTOMER CATEGORIES

Karolis Kiaunė, Simona Ramanauskaitė

Natural language processing and classification have been widely used in English-speaking countries. However, analysis and classification of a Lithuanian text is a complex issue and has not been fully implemented. This is due to complexity and peculiarities of the Lithuanian language, so methods appropriate for other languages, are not always appropriate for the Lithuanian language.

Three selected word processing options and their various combinations were used and it was assessed how different and consistent text classification methods are able to classify insurance company customers‘ enquiries sent by email. This study is unique because a great number of methods were used and classification accuracy of a Lithuanian text in a large number of categories (33) was further assessed.

Natural language processing problems, analogous studies of Lithuanian text classification were analyzed, research methodology was proposed and research findings were discussed in the paper.

Keywords: *NLP, text classification, emails, text processing*.

Santrauka

DIDELĮ KATEGORIJŲ KIEKĮ TURINČIŲ DRAUDIMO BENDROVĖS KLIENTŲ UŽKLAUSŲ, GAUTŲ ELEKTRONINIAIS LAIŠKAIS, LIETUVIŠKO TEKSTO KLASIFIKAVIMAS

Karolis Kiaunė, Simona Ramanauskaitė

Natūralios kalbos apdorojimas ir klasifikavimas jau plačiai naudojamas anglakalbėse šalyse. Tačiau lietuviško teksto analizė ir klasifikacija yra sudėtinga ir dar nevisiškai įgyvendinta. Taip yra dėl lietuvių kalbos sudėtingumo ir savitumo, todėl kitoms kalboms tinkami metodai ne visada tinka lietuvių kalbai.

Šiame straipsnyje naudojamos trys pasirinktos tekstų apdorojimo parinktys bei įvairios jų kombinacijos ir įvertinama, kaip skirtingi nuoseklūs teksto klasifikavimo būdai gali klasifikuoti draudimo bendrovės klientų užklausas, gautas el. paštu. Šis tyrimas yra išskirtinis naudojamų metodų gausa ir papildomai įvertina lietuviško teksto klasifikavimo tikslumą daugelyje (33) kategorijų.

Straipsnyje aptariamos natūralios kalbos apdorojimo problemos, analogiški tyrimai su lietuvių kalba parašytų tekstų klasifikacija, pristatoma siūloma tyrimo metodika ir aptariami tyrimo rezultatai.

Prasminiai žodžiai: *NLP, teksto klasifikavimas, elektroniniai laiškai, teksto apdorojimas.*

Įteikta 2019-10-26
Priimta 2019-11-21