Kauno technologijos universitetas
Matematikos ir gamtos mokslų fakultetas

Sentimento poliariškumo tyrimas vartotojų atsiliepimuose: žodynu ir mašininiu mokymu grįstų metodų palyginimas ir sujungimas

Baigiamasis magistro projektas

Diana Karosevičiūtė
Projekto autorius

Doc. dr. Evaldas Vaičiukynas
Vadovas
Doc. dr. Beata Šeinauskienė
Vadovas

Kaunas, 2018

Kauno technologijos universitetas
Matematikos ir gamtos mokslų fakultetas

Sentimento poliariškumo tyrimas vartotojų atsiliepimuose: žodynu ir mašininiu mokymu grįstų metodų palyginimas ir sujungimas

Baigiamasis magistro projektas
Didžiųjų verslo duomenų analitika (621G12002)

Diana Karosevičiūtė
Projekto autorius

Doc. dr. Evaldas Vaičiukynas
Vadovas
Doc. dr. Beata Šeinauskienė
Vadovas

Doc. dr. Kristina Šutienė
Recenzentas
Lekt. Linas Ablonskis
Recenzentas

Kaunas, 2018

Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Diana Karosevičiūtė

Sentimento poliariškumo tyrimas vartotojų atsiliepimuose: žodynu ir mašininiu mokymu grįstų metodų palyginimas ir sujungimas

Akademinio sąžiningumo deklaracija

Patvirtinu, kad mano, Dianos Karosevičiūtės, baigiamasis projektas tema „Sentimento poliariškumo tyrimas vartotojų atsiliepimuose: vektorių sudarymo ir klasifikavimo sprendimų palyginimas“ yra parašytas visiškai savarankiškai ir visi pateikti duomenys ar tyrimų rezultatai yra teisingi ir gauti sąžiningai. Šiame darbe nei viena dalis nėra plagijuota nuo jokių spausdintinių ar internetinių šaltinių, visos kitų šaltinių tiesioginės ir netiesioginės citatos nurodytos literatūros nuorodose. Įstatymų nenumatytų piniginių sumų už šį darbą niekam nesu mokėjęs.

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

Turinys

Paveikslų sąrašas

Lentelių sąrašas

Karosevičiūtė, Diana. Sentimento poliariškumo tyrimas vartotojų atsiliepimuose: žodynu ir mašininiu mokymu grįstų metodų palyginimas ir sujungimas. Magistro baigiamasis projektas / vadovai: doc. dr. Evaldas Vaičiukynas; doc. dr. Beata Šeinauskienė; Kauno technologijos universitetas, Matematikos ir gamtos mokslų fakultetas.

Studijų kryptis ir sritis (studijų krypčių grupė): Taikomoji matematika (A02), Matematikos mokslai (A).

Reikšminiai žodžiai: sentimentų analizė, teksto vektorizavimas, atpažinimo teorija, vartotojo pasitenkinimas, klientų atsiliepimai.

Kaunas, 2018. 62 p.

Santrauka

Sentimentų klasifikavimas vartotojų atsiliepimuose itin aktuali tema įmonėms, siekiančioms daryti išsamius tyrimus apie vartotojo patirtį ir pasitenkinimą. Automatinis sentimentų klasifikavimas leidžia einamuoju laiku stebėti klientų pasitenkinimo lygį, iš anksto reaguoti į svarbius pokyčius ir taip įgyti konkurencinį pranašumą.

Šiame darbe nagrinėjami ir lyginami vektorių iš teksto sudarymo metodai: PV-DBOW metodas derinant žodžių ir dokumentų vektorius, PV-DBOW metodas naudojant tik dokumentų vektorius, pastraipų vektorius – paskirstytos atminties modelis, latentinis semantinis indeksavimas, atsitiktinių projekcijų metodas ir Sent2Vec. Metodai buvo panaudoti sprendžiant sentimento poliariškumo detekcijos uždavinį su mašininio mokymo modeliais (logistinės regresijos, atsitiktinių miškų ir daugiasluoksnio perceptrono). Sentimento poliariškumas papildomai buvo įvertinamas naudojant žodynu grįstus metodus: SenticNet4, SentimentGI, SentimentHE, SentimentLM ir SentimentQDAP. Metodai buvo išmėginti su IMDB, TripAdvisor ir Amazon vartotojų atsiliepimų duomenų rinkiniais. Atrinkti geriausi skirtingų metodų rezultatai buvo sujungti sprendimų lygmenyje pritaikant atsitiktinio miško detektorių. Gautas hibridinis modelis pasiekė didžiausią klasifikavimo tikslumą lyginant su pavienių metodų taikymo rezultatais visiems duomenų rinkiniams.

Karosevičiūtė, Diana. Sentiment polarity detection in customer reviews: comparison and fusion of dictionary and machine learning based methods. Master's Final Degree Project / supervisors: assoc. prof. Evaldas Vaičiukynas; assoc. prof. Beata Šeinauskienė; Faculty of Mathematics and Natural Sciences, Kaunas University of Technology.

Study field and area (study field group): Applied mathematics (A02), Mathematics (A).

Keywords: sentiment analysis, text embedding, pattern recognition, customer satisfaction, customer reviews.

Kaunas, 2018. 62 pages.

Summary

Sentiment classification incustomer reviews is relevant for companies which seek to do research of its customer experience and satisfaction. Automatic sentiment classification empowers companies to track customer satisfaction in real time, take actions in advance and get competitive advantage.

Thesis work considers and compares vector embedding methods: Paragraph Vector – Distributed Memory Model (PV-DBOW) with document vectors and both document and word vectors, Latent Semantic Indexation, Random Projections and Sent2Vec. Customer reviews as input datasets are taken from Imdb, TripAdvisor and Amazon. Machine learning and dictionary based algorithms have been applied for data classification: Logistic Regression, Random Forests and Multilayer perceptron, SentimentGI, SentimentHE, SentimentLM, SentimentQDAP and SenticNet4 dictionary based algorithms. Six best modelling results were combined and random forest classification was applied. New hybrid model has got the highest classification AUC scores in comparison with methods before join for all datasets.

Įvadas

Žmonių bendravimas ir apsikeitimas nuomone – itin svarbi kasdienės žmogaus veiklos dalis ir yra vienas iš pagrindinių faktorių, nulemiančių individo elgesį. Žmonių įsitikinimai ir realybės suvokimas, priimami sprendimai yra didele dalimi sąlygoti aplinkinių mąstymo ir pasaulio vertinimo. Dėl šios priežasties, kai žmogui reikia priimti sprendimą, dažnai ieškoma kitų žmonių ir prašoma jų išreikšti savo nuomonę. Tai galioja ne tik individams, bet ir įmonėms. Sentimentai kaip ir nuomonė, įvertinimai, požiūriai ir emocijos – tai sentimentų analizės studijų objektas.

Augantis tyrėjų susidomėjimas sentimentų tematika sutampa su smarkiai padidėjusiu socialinės medijos populiarumu: vis daugiau žmonių išreiškia savo nuomonę atsiliepimų puslapiuose, žiniasklaidos publikacijų komentaruose, diskusijose, tinklaraščiuose ir socialiniuose tinkluose (pvz. *Twitter, Facebook*). Taip pirmą kartą istorijoje atsirado neįtikėtinai didelis duomenų, išreiškiančių kieno nors nuomonę, kiekis, prieinamas internetu. Sentimentų analizė yra viena iš labiausiai tyrinėjamų sričių visoje natūralios kalbos apdorojimo tematikoje, plačiai nagrinėjama ne tik duomenų ir teksto tyrybos srityse, bet ir sparčiai plinta iš informacinių technologijų srities į vadybos mokslus dėl savo svarbos verslui ir visuomenei. Pastaruoju metu suklestėjo ir industrinės veiklos, apimančios sentimentų analizę. Įsikūrė daug pradedančių įmonių, didesnės kompanijos pačios kuria sentimentų analizės įrankius savo poreikiams patenkinti. Dėl šios priežasties aktualu palyginti naujus ir klasikinius sentimentų analizės metodus, leidžiančius pačiai įmonei susikurti efektyviai veikiantį sentimento detekcijos įrankį.

Darbo objektas: sentimentai vartotojų atsiliepimuose.

Darbo tikslas: palyginti sentimento detekcijos metodus, grįstus mašininiu mokymu ir žodynu, ir ištirti jų sujungimo perspektyvumą.

Darbo uždaviniai:

1. teoriškai pagrįsti vartotojų sentimentų tyrimų aktualumą ir problematiką;
2. atlikti mokslinės literatūros analizę ir išsiaiškinti, kokie metodai naudojami sentimentų analizėje;
3. ištirti vektorių sudarymo ir sentimento detekcijos metodų, grįstų mašininiu mokymu ir žodynu, tinkamumą vartotojų atsiliepimų duomenų rinkiniams;
4. sudaryti hibridinį vartotojų sentimento detekcijos modelį, sujungiant pasiteisinusius metodus.

1. Literatūros apžvalga

Teksto analizės svarba išaugo kartu su interneto plėtra ir vartotojų skaičiaus augimu. Analizuojant tekstą gaunama aktuali informacija, kuri panaudojama daugybėje skirtingų sričių, pavyzdžiui, tyrėjai seka viešumoje populiarėjančias idėjas, naujienas, pastebimus didelius įvykius, tokius kaip rinkimai ar kiti politiniai įvykiai, akcijų rinkos svyravimai, prasidėjusios epidemijos ir atkreipia dėmesį į tuo metu visuomenei aktualiausias naujienas [1].

Tačiau didėjant informacijos kiekiui internete, vartotojui neužtenka laiko ir išteklių sekti visus informacijos šaltinius. Siekiant taupyti žmogiškuosius resursus ir automatizuoti informacijos gavimą, vis dažniau naudojami mašininio mokymo algoritmai.

Šiame skyriuje nagrinėjama sentimentų analizės samprata mokslinėje literatūroje, nauda ir poreikis vadybos srityje. Taip pat apžvelgiami atlikti naujausi tyrimai, mokslinės diskusijos, ir problemos, kylančios analizuojant vartotojų sentimentus ir jų poliariškumą vartotojų atsiliepimuose.

1.1.Sentimentų samprata ir analizė

Sentimentai yra nagrinėjami jau nuo 2001 metų [2], tačiau pirmieji darbai, kurių objektas yra sentimentų analizė, pasirodė apie 2003 metus [3]. Pagrindinis sentimentų analizės tikslas – nustatyti nuomonės apie objektą orientaciją, tai yra suklasifikuoti tekstinius duomenis į teigiamo, neigiamo ir neutralaus sentimento klases.

Pateikiamos sentimentų analizėje dažnai vartojamos sąvokos [4]:

- Sentimentas yra jausmas, požiūris, įvertinimas arba emocija, susieta su nuomone. Sentimentą galima išreikšti trimis savybėmis:

  $$(y, o, i) \quad (1)$$

  čia y – sentimento tipas; o – sentimento orientacija; i – sentimento intensyvumas.
- Sentimentų analizė, taip pat vadinama nuomonės tyryba (angl. *Opinion mining*) – tai mokslo šaka, nagrinėjanti žmonių nuomonę, sentimentus, įvertinimus, pagyrimus, požiūrius ir emocijas apie tam tikrus objektus, pavyzdžiui, paslaugas, prekes, žmones, kompanijas, problemas, įvykius ir jų atributus [5]. Literatūroje randama daug terminų, kurie nusako sentimentų analizėje nagrinėjamos problemos sritį: atsiliepimų tyryba (angl. *review mining*), emocijų analizė (angl. *emotion analysis*), afekto analizė(angl. *affect analysis*), sentimentų analizė, nuomonės tyryba.
- Sentimento orientacija – sentimento savybė turėti teigiamą, neigiamą arba neutralią emociją. Sentimento orientacija literatūroje dar vadinama poliariškumu.
- Sentimento intensyvumas išreiškiamas žodžiais, stipriai arba silpnai išreiškiančiais nuomonės reiškėjo sentimentą. Pavyzdžiui, pirkinys gali būti apibūdintas kaip *geras*, tačiau

pasakydamas, kad pirkinys buvo *nuostabus*, vartotojas daug stipriau išreiškia savo teigiamą sentimentą. Taip pat gali būti vartojami sentimentą sustiprinantys arba silpninantys žodžiai, tokie kaip *labai, itin, neįtikėtinai, siaubingai.*

- Sentimento reitingas gali būti išreikštas skaitine reikšme, parodančia sentimento intensyvumą, pavyzdžiui, skalėje nuo 1 iki 5 žvaigždučių. Praktikoje dažniausiai naudojama skalė iki penkių, nes suklasifikuoti sentimentus į daugiau klasių algoritmui gali būti sudėtinga, dėl to labai maži skirtumai nulemia kitokį rezultatą. Tokią skalę galima interpretuoti vartotojui suprantamomis kategorijomis:
    1. emocionalus teigiamas (5 žvaigždutės);
    2. racionalus teigiamas (4 žvaigždutės);
    3. neutralus (3 žvaigždutės);
    4. racionalus neigiamas (2 žvaigždutės);
    5. emocionalus neigiamas (1 žvaigždutė).

Vartotojo dažnai prašoma reitingu įvertinti produktą ar paslaugą, nes tada galima lengviau suklasifikuoti nuomones. Šie reitingai taip pat labai naudingi apmokant mašininio mokymo algoritmus, nes nebūtina naudoti žodynų, kurie priskiria tam tikriems žodžiams dažniausiai pasitaikančius sentimentus. Tačiau tada kyla problemos, kaip įvertinti semantinę nuomonės prasmę.

Tekste gali būti išreiškiama nuomonė, kurią ir analizuoja sentimentų analizė, arba faktinė informacija. Bendrai teksto analitika nagrinėja abi teksto rūšis, o dažniausias faktinės informacijos panaudojimas yra paieška internete, pagrįsta faktais ir objektų tarpusavio sąsajomis. Nuomonė gali būti išreikšta žymėjimu [4]:

$$(e, a, s, h, t) \tag{2}$$

čia e – objektas; a – objekto savybė, apie kurią reiškiama nuomonė; s – nuomonės sentimentas (teigiamas, neigiamas, neutralaus arba kiekybinis įvertis); h – nuomonės reiškėjas; t – nuomonės išreiškimo laikas.

Visos penkios komponentės labai svarbios. Tuo atveju, kai objektas apibūdinamas kaip visuma, galima pašalinti a, t. y., objekto savybę, tačiau įmonės teikiamų produktų ar paslaugų, susidedančių iš dalių, nuomonės išskyrimas pagal savybes gali būti kritiškai svarbus. Pavyzdžiui, telefono kamera gali būti itin giriama už kokybę, tačiau atsparumas vandeniui labai prastas. Abu aspektus sudėjus į vieną komentarą, algoritmas gali nuomonę suklasifikuoti neteisingai, o pardavėjas neteks svarbios informacijos apie produktą ar paslaugą.

Identifikuoti nuomonės reiškėją gali būti svarbu tuo atveju, kai nuomonę išreiškia ne eilinis žmogus, o visuomenei žinoma asmenybė arba neigiamą nuomonę išreiškia nuolatinis klientas. Nuomonės išreiškimo laikas gali padėti sekti nuomonės kitimo laike tendenciją.

- Sentimento tipas gali būti priklausomas nuo pasirinktos vertinimo srities, pavyzdžiui, pagrįstas lingvistika, psichologija arba vartotojų tyrimu. Vartotojų tyrimai klasifikuoja sentimentus į racionalius ir emocinius. Racionalūs – pagrįsti argumentavimu, objektyviomis priežastimis ir logika. Emocionalūs sentimentai yra subjektyviai argumentuojami, stipriai priklauso nuo nuomonės reiškėjo psichologinės būsenos [4].

Siekiant vykdyti paveikias rinkodaros kampanijas, iškeliami tikslai sukelti vartotojams ne tik teigiamus, bet ir emociškai stiprius sentimentus.

Semantinė analizė gali būti vykdoma 3 skirtingais lygmenimis: žodžio, sakinio ir dokumento. Sentimentų tyrėjai susiduria su problemomis, kai neaišku, kaip pasirinkti, kokiu lygmeniu analizuoti tekstą. Taip pat reikia įvertinti kalbos subtilumą: vartotojas gali naudoti ironiją arba sentimentą, išreikštą neutraliais žodžiais. Tie patys žodžiai skirtingose veiklos srityse turi kintančią prasmę, todėl svarbu analizuoti kontekstą. Semantinei nuomonės prasmei didelę reikšmę turi ne tik žodžiai, bet ir sintaksė.

Apibendrinant, sentimentų analizė yra plati teksto analizės sritis, nagrinėjanti nuomonės apie objektą orientaciją, kuri dažniausiai skirstoma į teigiamą, neigiamą ir neutralią klases. Analizuojant sentimentus, svarbu atkreipti dėmesį į jų tipą, orientaciją, intensyvumą ir reitingą. Reitingai itin padeda naudojant mašininio mokymo algoritmus, nes nereikia naudoti sentimentų žodynų. Dar vienas analizės objektas yra nuomonė, sudaryta iš nagrinėjamo objekto ar jo savybės, sentimento, nuomonės reiškėjo ir laiko, iš kurių kiekvienas suteikia svarbios informacijos. Vykdant sentimentų analizę svarbu išsiaiškinti, kokiu lygmeniu verta analizuoti tekstą, įvertinti iššūkius: kalbos subtilumą, daugiareikšmiškumą, sintaksę, semantiką.

1.2.Sentimentų analizė rinkodaroje

Metodai, kurie leidžia nustatyti teigiamas arba neigiamas vartotojo nuostatas apie specifinį subjektą, tokį kaip įmonė arba jos produktai, turint didelį dokumentų kiekį, gali būti pritaikomi įvairiose srityse. Tai atveria galimybes efektyviau valdyti riziką, keliamą negatyvių gandų apie produktą arba įmonės vardą. Pavyzdžiui, Emil‘is ir Salib‘as [6] nagrinėjo JAV oro bendrovės „United Airlines“ krizę 2017 metais. Paaiškėjo, kad staigiai paplitęs neigiamas sentimentas apie grubų elgesį su keleiviais turėjo trumpalaikį neigiamą efektą kompanijos rinkos vertei, tačiau nenulėmė vertės ilguoju laikotarpiu.

Taip pat sentimentų analizė leidžia pagerinti įmonės rinkodaros ir konkurencingumo analizę. Mokslininkai S. Chan ir M. Chong atliko sentimentų analizę su finansų srities tekstais ir gavo statistiškai reikšmingus rezultatus, atskleidžiančius, jog tekstuose išreiškiami sentimentai padeda prognozuoti akcijų rinkos indeksų vertes, nors įprastai tokie sentimentai ir akcijų rinkos indeksai yra

laikomi nekoreliuotais [8]. Pastaraisiais metais daugėja autorių, tiriančių sentimentų poveikį akcijų rinkoms [9][10][11].

Dažniausi sentimentų analizės sprendžiami uždaviniai: atsiliepimų klasifikavimas į teigiamus ir neigiamus, produkto atsiliepimų nagrinėjimas, tam tikros temos sentimentų kitimo tendencijų laike sekimas, rinkimų arba rinkos prognozė ateičiai.

Priklausomai nuo nagrinėjamos veiklos srities sentimentų analizė gali būti pritaikoma įvairaus pobūdžio vartotojų kuriamam atsiliepimų ir nuomonės formavimo srautui internete [1]:

- klausimų – atsakymų duomenų bazės (Yahoo Answers, Ask.com ir kt.);
- video medžiaga internete (Youtube, Vimeo ir kt.);
- asmeniniai tinklaraščiai (Blogger, Weebly ir kt.);
- mikrotinklaraščiai (angl. *microblog*) (Tumblr, Twitter ir kt.);
- muzikos įrašų tinklaraščiai (iTunes ir kt.);
- atsiliepimų tinklapiai (Yelp, TripAdvisor ir kt.);
- socialiniai tinklai (Facebook, MySpace ir kt.);
- enciklopedijos, žinynai internete (Wikepedia).

Neefektyvios įmonės daug pinigų išleidžia tradicinėms rinkodaros veikloms, pavyzdžiui, darbuotojai rengia klientų pasitenkinimo apklausas ir vėliau analizuoja rezultatus rankiniu būdu arba naudodami paprastas programas. Tačiau investavę į technologijas, kompanijos tiesiogiai naudoja iš sentimentų analizės gaunamus duomenis savo produkto priimtinumo ir įvaizdžio rinkoje sekimui ir prognozavimui. Dabartinės technologijos leidžia naudoti sentimentų analizę netgi realiu laiku.

Pastaraisiais metais daug tyrimų atliekama pritaikant sentimentų analizę vartotojų atsiliepimams nagrinėti įvairiose rinkose. Mokslinių tyrimų rezultatų sentimentų analizės rinkodaros tematikoje apibendrinimas pateiktas 1 lentelėje.

1 lentelė. Mokslinių tyrimų rezultatų sentimentų analizės rinkodaros tematikoje apibendrinimas

Literatūros šaltinis	Veiklos sritis / industrija	Algoritmas	Tikslas	Duomenų šaltinis	Rezultatai	Apribojimai / tolimesnių tyrimų kryptys
WU, J., 2017 [12]	Internetinė prekyba	Veiksnių hierarchijos modelis (angl. *Hierarchy-of-effects model*)	Išsiaiškinti vartotojų atsiliepimų efektyvumą lemiančius veiksnius, kurie padėtų priimti rinkodaros valdymo sprendimus	Amazon atsiliepi-mai apie du produktus	Sukurtas empiriškai pagrįstas modelis, kuris atskleidė, kad atsiliepimų išsamumas yra svarbus naudingumo veiksnys pirkėjams ir pardavėjams	Rezultatų generalizavimui turėtų būti panaudota daugiau ir didesnių duomenų rinkinių
KARIMI, S. - WANG, F., 2017 [13]	Internetinė prekyba	Gradientinis optimizavimo algoritmas (angl. *gradient boosting algo-*	Padėti vartotojams kurti naudingesnius atsiliepimus ir	Amazon atsiliepima i apie tris produktų grupes	Sukurtas modelis, kuris naudojasi atsiliepimo poliariškumu, subjektyvumu,	Nenagrinėjami kita kalba nei anglų parašyti žodžiai, analizuojami tik

		rithm)	patobulinti įmonių tinklapius naudojantis naudingiausių atsiliepimų reitingavimo sistema		entropija ir skaitymo sudėtingumu, prognozuoja naudingumą vartotojui bei pranoksta tiesinės regresijos rezultatą	viename tinklapyje esantys duomenys, ateityje gali būti įtraukta daugiau reikšmingų kintamųjų
ETTER, M. et al., 2018 [14]	Bankų sektorius	Mašininis mokymas, pasyvus – agresyvus klasifikatorius (angl. *Passive–Aggressive (PA) classifier*)	Pasiūlyti metodą pamatuoti organizacijos legitimumą ir visuomenės pasitikėjimą organizacija naudojantis socialine žiniasklaida ir sentimentų analize	14,179 įrašai iš Twitter tinklapio ir 722 straipsniai apie Italijos banką	Iš Twitter ir straipsnių analizės gauti rezultatai nesutampa, todėl siūloma įmonės padėčiai įvairiapusiškai analizuoti naudojant socialinių tinklų ir straipsnių analizę	Atsiliepimai nepatenka į analizę, jei vartotojas Twitter paskyroje tiesiogiai nemini banko. Viešųjų ryšių įmonių paskleisti komentarai gali iškreipti analizės rezultatus
ALAEI, A.R. et al., 2017 [15]	Turizmo sektorius	Mašininis mokymas, leksikonu grįstas metodas, artimiausio kaimyno metodas ir jų patobulinimai	Palyginti kelių metodų veikimą pasitelkiant turizmo sektoriaus duomenis	Sanderso Twitter duomenų rinkinys	Levallois metodas geriau veikia binariniu atveju, o VADER metodas – trijų klasių atveju. Metodai geriausiai klasifikuoja teigiamus komentarus.	Reikia labiau pritaikytų žodynų, skirtų turizmo sektoriui, neiginių ir neutralių komentarų atpažinimas yra semantiškai sudėtingos užduotys
XU, Z. et al., 2011 [16]	Ekstrema-lių įvykių valdymas	Dažniausiai pasikartojan-tys žodžiai	Pagerinti ekstremalių įvykių valdymo sistemų efektyvumą ir tikslumą	3321 Sina Weibo mikrotink-laraščio įrašai apie taifūną Chan-hom	Sukurtas dalyvių pojūčiais grįstas modelis gauti esamos padėties informaciją ekstremalių įvykių atveju	Komentare turi būti pažymėta asmens lokacija, todėl tinkamų komentarų skaičius nedidelis
WEGBA, K. et al., 2017 [17]	Filmų rekomen-dacijos	Latentinė semantinė analizė (LSA)	Pasiūlyti interaktyvią rekomendavim o sistemą	MovieLens 100K duomenų bazės atsiliepima i apie filmus	Filmų rekomendavimo sistema sujungiant sentimentų analize grįstą rekomendaciją ir istorijų pasakojimą	Trūksta formalaus rekomendavimo sistemos efektyvumo vertinimo, didesnės rekomendavimo istorijų įvairovės

Iš nagrinėtų mokslinės literatūros šaltinių matoma, kad sentimentų analizė yra naudingas įrankis rinkodaroje, teikiantis naudą tiek vartotojui, tiek produkto ar paslaugos pardavėjui. Naudojant sentimentų analizę galima atrinkti daugiausiai informacijos suteikiančius vartotojų atsiliepimus klientams, rekomenduoti aktualius ir patinkančius filmus ar kitus produktus, sutaupyti klientų laiko ieškant tinkamiausio produkto. Įmonės taip pat laimi išaugusius pardavimus, padidėjusį klientų pasitenkinimą paslaugomis, geresnę vartotojo patirtį. Taip pat gali sekti realiu laiku savo organizacijos legitimumą ir visuomenės pasitikėjimą analizuojant informaciją žiniasklaidoje ir socialiniuose tinkluose bei greičiau reaguoti į kylantį vartotojų nepasitenkinimą. Šių technologijų

vystymas turi didelę naudą ne tik komercinei veiklai, bet ir valstybinėms įmonėms, nes įgalina stebėti viešosios nuomonės pokyčius visuomenėje ir automatizuoti virtualios erdvės stebėjimą.

Apibendrinus, dažnai minimi tyrimų apribojimai yra nedidelis duomenų kiekis iš vieno sektoriaus, tačiau norint generalizuoti tyrimo rezultatus reiktų atlikti tyrimus su daugiau duomenų iš įvairių sričių. Taip pat, jeigu naudojami leksikonu grįsti metodai, kyla kalbos apribojimų, nes dauguma žodynų sudaryti anglų kalba ir pagrinde skirti tik konkrečiai sričiai. Ateityje duomenų kiekiai dar labiau išaugs, todėl ir tyrimai turi būti atliekami su didesniais ir įvairesniais duomenimis. Taip pat, reikia atkreipti dėmesį į netiesioginę arba perkeltinę prasmę turinčius atsiliepimus, kuriuose gali būti užslėptas tiek sentimentas, tiek objektas, apie kurį išreiškiama nuomonė.

Rinkoje galima rasti keletą komercinių ir akademinių įrankių, skirtų sekti viešąją nuomonę dideliu mastu, pavyzdžiui, *IBM, Oracle, SAS, SenticNet* ir kiti. Dauguma leidžia įrankiais naudotis žmonėms, neturintiems aukštos kompetencijos technologijose ir siūlo grafinį apibendrinimą. Pagrindiniai trūkumai yra ribotas kalbos pasirinkimas, nes dažniausiai leidžiama tik anglų kalba. Taip pat, jie fiksuoja labai aiškiai išsakomą nuomonę, tačiau neteisingai nustato užslėptą arba ironiškai išreikštą sentimentą.

1.3.Sentimentų analize grįsti vartotojų pasitenkinimo tyrimai

Vartotojo pasitenkinimas parodo, ar kompanijos teikiami produktai ir paslaugos atitinka arba viršija kliento lūkesčius. Pasitenkinimas gali būti matuojamas kaip procentinė dalis klientų, patenkintų kompanija, jos teikiamomis paslaugomis, produktais arba kaip dalis klientų, kurių išmatuotas pasitenkinimo lygis viršija iš anksto įmonės užsibrėžtą minimalų klientų pasitenkinimo lygį. Tai svarbi rinkodaros strategijos dalis. Atlikus patyrusių rinkodaros vadovų apklausą, daugiau nei 71% respondentų teigė, kad vartotojo pasitenkinimo lygio matavimas leido efektyviau stebėti ir valdyti verslą [18].

Vartotojo pasitenkinimas tradiciniais metodais dažnai matuojamas klientų pasitenkinimo apklausomis. Šių apklausų rezultatai atneša pakankamai nedidelę naudą dėl šių trūkumų:

- fiziškai galima apklausti tik nedidelę klientų imtį, kuri ne visada būna reprezentatyvi;
- sudėtinga tinkamai sudaryti klausimynus, kurie atskleistų visą informaciją, kuri domina paslaugų ar prekių teikėją ir leistų sužinoti specifinio kliento nuomonę

Dėl šių trūkumų, užuot tyrusios kliento pasitenkinimą apklausomis, vis daugiau įmonių nori sužinoti vartotojo pasitenkinimą internete esančiuose šaltiniuose: savo tinklapyje, specifiniuose produktų apžvalgų tinklapiuose, forumuose ir pastaraisiais metais itin daug dėmesio sulaukiančiuose socialiniuose tinkluose. D.Kang‘as ir Y. Park‘as, teigia, kad tradicinių vartotojų pasitenkinimo apklausų rengimas užima daug laiko, tad pasiūlė metodologiją, kaip matuoti vartotojų pasitenkinimo lygį pusiau automatiškai, naudojant sentimentų analizę. Anot autorių, pasiūlytas požiūris turi panašų

efektyvumą lyginant su tiesioginėmis vartotojų apklausomis ir gali būti svarbus pagrindas tolimesniems vartotojų pasitenkinimo tyrimams [19].

Jei kaštų sumažinimo ir pelno maksimizavimo sprendimai priimami vietoj investavimo į paslaugų gerinimą ir klientų pasitenkinimą, tai yra trumparegiškas požiūris. Daugėja įrodymų, kad investicijos į klientų pasitenkinimą atsiperka ilguoju laikotarpiu, tačiau tam reikia turėti objektyvių matų, kurie padėtų stebėti pažangą. Grįžtamasis ryšys iš personalo, tiesiogiai kontaktuojančio su klientais, gali būti subjektyvus, o skundų tyrimas nėra reprezentatyvus [20]. Tokiu atveju gali būti pasitelkiami vartotojų atsiliepimai internete.

Žmogus gali nesudėtingai nustatyti, kokią emociją savo komentaru perduoda paslaugų ar produkto vartotojas, tačiau sekti visą informacijos srautą apie įmonės vardą skirtinguose šaltiniuose pernelyg brangu. Visgi stebėti atsiliepimus internete yra itin svarbu, kadangi jie formuoja viešąją nuomonę apie įmonę, o paskleisti neigiami gandai gali sugadinti įmonės reputaciją ir sukelti komunikacijos krizę.

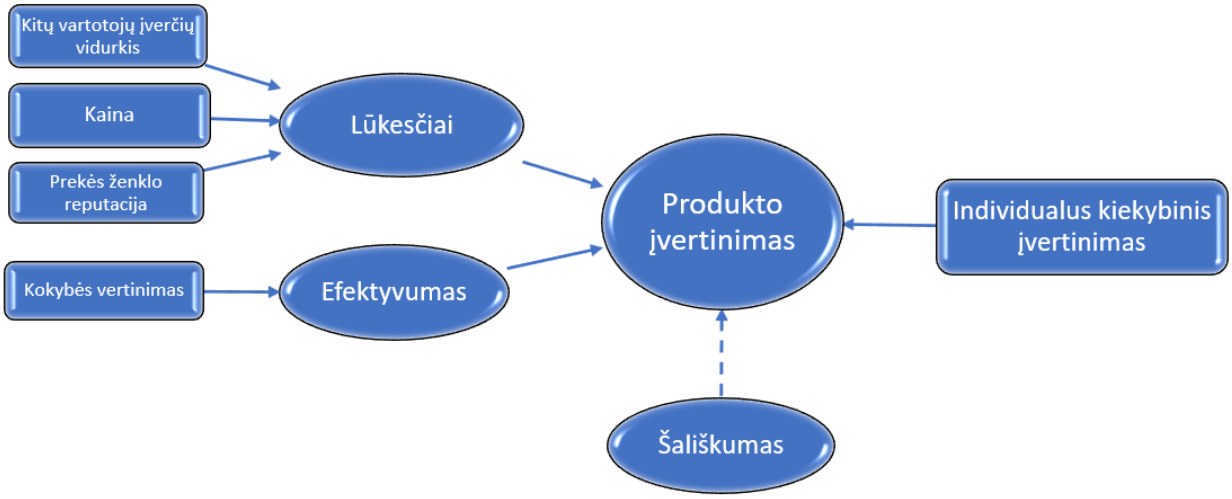
Socialinė žiniasklaida pakeitė bendravimą tarp pardavėjų ir pirkėjų įgalindama pardavėjus įsitraukti į bendravimą socialiniuose tinkluose ir taip pasiekti geresnius pardavimų rezultatus. Aktyviai atliekami tyrimai analizuojantys sąryšį tarp aktyvios pardavėjų komunikacijos socialinėje žiniasklaidoje ir klientų pasitenkinimo iš vertės kūrimo perspektyvos. Tyrimų rezultatai atskleidžia, kad socialinės žiniasklaidos naudojimas pagreitina pardavėjų reakciją į vartotojų atsiliepimus ir padidina klientų pasitenkinimą verslas verslui (angl. *business-to-business, B2B*) srityje [21].

A. Sengupta tyrinėjo skirtingas vartotojų reakcijas susidūrus su nekokybiškomis paslaugomis [22]. Nustatyta, kad vartotojų reagavimo būdai skiriasi priklausomai nuo to, kaip stipriai nepatenkinamai buvo suteiktos paslaugos. Taip pat išsiaiškinta, kad esant tam tikroms sąlygoms, teigiama prekės ženklo reputacija gali sušvelninti neigiamą vartotojų reakciją ir paveikti vartotojų elgseną netgi pastebėjus pakankamai didelius trūkumus ir nekokybiškas paslaugas.

Vartotojo pasitenkinimo sustiprinimas gali atnešti daug naudos. Įvairiose industrijose patenkinti klientai išleidžia daugiau pinigų ir pasižymi didesniu lojalumu [20]. Pavyzdžiui, bankininkystės sektoriuje septynis kartus labiau tikėtina, kad klientai padidins savo depozitą, ir du kartus labiau tikėtina, kad atsidarys papildomą sąskaitą, jei jie vertina banko paslaugas kaip puikias (vartotojo pasitenkinimo rodiklis yra devyni arba dešimt dešimties balų sistemoje) lyginant su vidutiniu įvertinimu (nuo šešių iki aštuonių iš dešimties). Panašiai, mokamos televizijos klientai, kurie įvertina tiekėją puikiai, tikėtina, gali du kartus ilgiau pasilikti su tuo pačiu tiekėju nei klientai, kurie įvertina vidutiniškai arba prasčiau [23].

T.H. Engler‘is su kolegomis sudarė vartotojų pasitenkinimo modelį internete esantiems produktų įvertinimams. Šis modelis kaip pagrindinius elementus ima prieš produkto ar paslaugos įsigijimą turimus lūkesčius ir po įsigijimo gautą rezultatą. Prieš produkto ar paslaugos įsigijimą

turimus lūkesčius suformuoja kitų vartotojų atsiliepimai, kaina ir prekės ar paslaugos ženklo reputacija. Po įsigijimo vartotojas patikrina kokybę ir taip nusprendžia apie prekės ar paslaugos veiksmingumą. Pridėjus kiekybinį reitingą (dažnai vertinama 5 žvaigždutėmis) ir išankstinį šališką kliento nusistatymą kompanijos ar produkto atžvilgiu, gaunamas bendras produkto įvertinimas [24].



1 pav. Vartotojo pasitenkinimo modelis internete esantiems produktų įvertinimams

Technologijų pažangos poveikį prekybos tinklų verslo modeliams nagrinėjo ir U. Ramanathan‘as su kolegomis, mokslinėje literatūroje pristatę tyrimą apie naujus verslo modelius, skirtus išlaikyti klientus ir įgyti konkurencinį pranašumą. Anot mokslininkų, prekybos tinklai turėtų analizuoti socialinėje žiniasklaidoje esančius klientų atsiliepimus, nes jie turi stiprų poveikį vartotojo pasitenkinimui. Socialinėje žiniasklaidoje esančių vartotojų atsiliepimų analizė leidžia plėtoti naujus verslo modelius, pasižyminčius unikalių paslaugų pasiūla ir inovatyviu rinkodaros požiūriu, didinančiu vartotojų lojalumą ir kuriančiu vertę klientams. Priešingai, autoriai nerado tvirtų argumentų, pagrindžiančių teigiamo ar neigiamo pasitenkinimo prekės ženklu poveikį vartotojų pasitenkinimui. Tačiau minėto tyrimo rezultatai parodė, kad vartotojo pasitenkinimui didelę įtaką daro akcijų vertės ir aptarnavimo operacijų sąveika [25].

Tiriant penkių aptarnavimo kokybės dimensijų poveikį vartotojų pasitenkinimui Jungtinės Karalystės greitojo maisto rinkoje buvo identifikuoti didžiausią įtaką turintys veiksniai [26]. Analizės rezultatai parodė, kad fizinis pokytis, reagavimas ir užtikrinimas yra svarbiausi veiksniai, po kurių seka patikimumas ir empatija.

Anot C. Fornell‘o ir jo kolegų, vartotojo pasitenkinimas taip pat yra reikšmingas kintamasis darant investicinius sprendimus akcijų rinkoje. Ištirta, kad aukštesnis vartotojų pasitenkinimas lemia didesnę kompanijos apyvartą ir gerokai rinkos vidurkį viršijančius dividendus. Šie rezultatai gauti įvertinus kitus alternatyvius paaiškinimus besiremiančius kompanijų dydžiu, verte, rizikos faktoriais. Taip pat įvertinta veikla skirtinguose technologijų ir pramonės sektoriuose lyginant NASDAQ ir DJIA akcijų indeksus [27].

Vartotojo pasitenkinimo efektas itin pastebimas vertinant bendrą akcininkų grąžą. Lyginant bendrą akcininkų grąžą tarp kompanijų, turinčių aukštesnę arba žemesnę vartotojo pasitenkinimo rodiklio reikšmę nei vidurkis, lyderiai pasiekia keturis kartus didesnį įmonės vertės augimą nei atsiliekančios kompanijos per 10 metų laikotarpį [23].

Įmonės, turinčios ilgalaikę perspektyvą, investuoja žmogiškuosius ir materialius išteklius į vartotojų pasitenkinimo lygio matavimą ir didinimą. Mokslinėje literatūroje pažymima, kad įvairiose industrijose patenkinti klientai išleidžia daugiau pinigų ir rodo didesnį lojalumą kompanijai, todėl įmonės pasiekia geresnius veiklos rezultatus pardavimo apimčių atžvilgiu, investuotojai daugiau investuoja akcijų rinkoje, kompanija įgyja konkurencinį pranašumą. Kadangi įprastinės vartotojų pasitenkinimo apklausos turi trūkumų, tokių kaip reprezentatyvumo arba objektyvumo stoka, didelės laiko ir finansų sąnaudos, organizacijos ieško kitų būdų kaip vartotojų pasitenkinimą įvertinti kiekybiškai. Šiai užduočiai spręsti siūloma analizuoti vartotojų atsiliepimus internete naudojant sentimentų analizę, kas įgalina sekti vartotojų pasitenkinimą realiu laiku.

1.4.Vartotojų patirties identifikavimas sentimentų analizės pagalba

Tyrėjai V. Alderson‘as ir L. Abbott‘as savo darbuose vieni pirmųjų didelį dėmesį skyrė idėjai, kad žmonės iš tikrųjų nori ne produkto, o patenkinančios patirties [28][29]. Vėlesniuose darbuose žmogaus elgesys buvo tyrinėjamas teikiant didelę svarbą sprendimų priėmimo ir patirties emociniams aspektams. B. J. Pinas ir J. G. Gilmore‘as atskyrė vartotojo patirtį nuo įsigyjamų produktų ar paslaugų pabrėždami, kad klientas įsigyja patirtį tam, kad maloniai leistų laiką bendraudamas ir gaudamas potyrius, kuriuos suteikia įmonė jam pritaikytu asmenišku būdu [30].

B. Schmitt‘as, L. Brakus‘as ir L. Zaratonello‘as teigia, kad bet kokios formos ir prigimties apsikeitimas paslaugomis ar produktais veda prie vartotojo patirties [31]. Kiti autoriai vartotojo patirtį vertina kaip holistinę, apimančią kartu vartotojo pažintinį, emocinį, jutiminį, socialinį ir dvasinį atsaką į visas sąveikas su įmone. Išskiriami keturi skirtingi požiūriai į vartotojų patirtį pagal tai, kiek dėmesio skiriama fiziniams ir emociniams paslaugos ar produkto komponentams ir pagal kokybinį arba humanistinį požiūrį į paslaugų teikimą (2 lent.) [32]. Taip pat teigiama, kad šiuo metu dauguma įmonių vadovaujasi funkciniu ir kokybiniu požiūriu, todėl daugiau dėmesio kreipia į formalų paslaugų ar produkto funkcionavimą, t.y. kad būtų patenkinti kliento poreikiai. Humanistinis požiūris labiau orientuojasi žmogų kaip pagrindinį prioritetą ir teikia į jį orientuotą produktą ir paslaugų dizainą.

2 lentelė. Skirtingi požiūriai į vartotojo patirtį

	Pabrėžiami funkciniai komponentai, kurie turi didžiausią įtaką vartotojo patirčiai	**Pabrėžiami emociniai komponentai, kurie teigiamai sustiprina vartotojo patirtį**
Sutartyje apibrėžtas ir kokybinis požiūris į paslaugų dizainą	Dabartinis požiūris: paslaugų kokybės struktūra	2 požiūris: emocionaliai įtraukianti patirtis
Humanistinis požiūris į paslaugų dizainą ir paslaugų teikimą	1 požiūris: humanistinis požiūris į dizainą ir paslaugų teikimą	3 požiūris: humanistinis požiūris į dizainą ir paslaugų teikimą su emocionaliai įtraukiančia patirtimi

Šaltinis: R. Bolton *et al.* [32]

Anot C. Meyer‘io ir A. Schwager‘io, verslo praktikoje vartotojo patirtis apibrėžiama kaip apibendrinanti visus kompanijos pasiūlymus ir paslaugas klientui: vartotojo aptarnavimo kokybę, reklamą, pakavimą, produkto ir paslaugų funkcionalumą, panaudojimo paprastumą ir patikimumą. Tai yra vidinis ir subjektyvus vartotojo atsakas į tiesioginį ir netiesioginį kontaktą su kompanija [33]. Teksto analitika ir mašininis mokymas gali padėti surasti aktualius vartotojų atsiliepimus internete ir jais remiantis atskleisti produkto ar paslaugos vartotojų patirtį. S. Hedegaard‘o ir J.G. Simonsen‘o atliktame tyrime paaiškėjo, kad 13%-49% komentarų iš jų nagrinėtų 3492 klientų atsiliepimų apie programinę įrangą ir video žaidimus turėjo informacijos apie vartotojo patirtį. Tyrimo rezultatai atskleidė, kad vartotojų atsiliepimų analizė gali suteikti organizacijai naudingų įžvalgų apie skirtingas produkto tinkamumo (angl. *usability*) ir vartotojo patirties dimensijas [34].

A. Kranzbühler‘is su kolegomis nagrinėjo dvilypę vartotojo patirties tyrimų prigimtį: literatūra statinę vartotojo patirtį nagrinėja viename konkrečiame laiko taške, tačiau tyrimai apie dinaminę patirtį atsižvelgia į vartotojo patirties kitimą ilgesniame laiko periode. Abi teorinės perspektyvos nagrinėja tą patį fenomeną – vartotojo patirties kūrimas iš organizacijos perspektyvos ir suvokimas iš vartotojo perspektyvos. Anot autorių, yra didelis potencialas produktyviai abiejų požiūrių simbiozei, todėl buvo pasiūlyta abiejų perspektyvų įžvalgų integracija [35].

Autorius Mccoll-Kennedy‘is sutinka, kad vartotojo patirtis turėtų būti analizuojama dinamiškai visos vartotojo kelionės (angl. *customer journey*) metu ir įvairiais aspektais, todėl pasiūlė tyrėjams analizuoti, kaip patirtis keičiasi per paslaugų ar produkto gyvavimo ciklą, paties vartotojo patirties kitimą. Išsamią analizę pasiūlyta atlikti trimis lygiais [36]:

1. retrospektyviai kaip detektyvui rekonstruoti vartotojo kelionę ir įvykius, atskleidžiančius motyvus, priemones ir galimybes;
2. realiu laiku analizuojant vykstančius įvykius ir vartotojo ir organizacijos sąveiką;
3. ekstrapoliuojant ateitį (simuliacijos, modeliavimo metodai).

Vartotojai, aktyviai skatinami organizacijos, gali įsitraukti į skirtingas veiklas ir kartu kurti vertę ypač pabrėžiant bendravimo ir tarpusavio sąveikos svarbą, įvairių vaidmenų kuriant paslaugų

patirtį, pavyzdžiui, reprezentavimo, apsikeitimo nuomone, patirtimi [37]. Kadangi vartotojo patirtis yra dinamiškas procesas, jis stipriai priklausomas nuo to, ką klientai daro ir sako, t.y. individai daro įtaką vienas kitam ir susidaro nuolat besikeičianti įtakos ekosistema [36]. Dėl to organizacijoms itin aktualu tirti šią ekosistemą realiu laiku.

K. Bauman‘as, B. Liu, ir A. Tuzhilin‘as pasiūlė naują rekomendavimo klientams sistemą naudojantis sentimentų analize vartotojų atsiliepimuose, kuria klientams siūlomas ne tik dominantis produktas, bet ir sąlygos, pagerinančios vartotojo patirtį. Siūlomas metodas prognozuoja sentimentus, kuriuos vartotojas gali patirti dėl įvairių produkto ar paslaugos aspektų ir identifikuoja vertingiausius aspektus, kurie gali potencialiai sukelti gerus sentimentus vartotojui [38].

Anot S. Chheda‘os, E. Dunkan‘o, S. Roggenhofer‘io skaitmenizavimas keičia vartotojo patirtį kiekviename sektoriuje, teikdamas naujų radikalių veiklos pasiūlymų, sutrikdydamas įprastus bendravimo tarp kompanijų ir klientų būdus ir keldamas itin aukštus reikalavimus paprastumui, personalizavimui ir interaktyvumui. Prieš pradedant analizuoti vartotojo patirtį reikia atkreipti dėmesį į savo produktą, kainą, paslaugą ir prekės ženklą. Jei produktas nepatikimas arba kaina per aukšta, net pati maloniausia vartotojo patirtis to nekompensuos. Kai esminiai dalykai sutvarkomi, kompanija gali išrinkti esmines vartotojo keliones, kurios labiausiai rūpi klientui, įvertinti savo veiklos atsiperkamumą, kad galėtų teisingai sudėti prioritetus ir gauti didžiausią grąžą iš investicijų [23].

Staigus išmaniųjų technologijų plitimas ne tik jaunimo tarpe, bet ir vyresnių vartotojų tarpe tampa įprastu reiškiniu ir suteikia galimybę iš organizacijų reikalauti vis efektyvesnių paslaugų. Kompanijos jau priėmė vartotojo patirties valdymo koncepciją ir supranta išmaniųjų įrenginių teikiamas galimybes prekybos srityje bei pradėjo jas tyrinėti [39]. Puikios vartotojo patirties kūrimas yra pagrindinis tikslas mažmeninės prekybos srityje tiek tradicinės prekybos būdu [40], tiek prekybos internetu [41].

McKinsey ištyrė bankininkystės sektoriaus kompanijų veiklą ir pateikia įmonių vartotojo patirties ir kelionės tobulinimo žingsnius, kuriuos reikia atlikti siekiant gauti didesnę vertę iš skaitmenizavimo proceso [23]:

- išsiaiškinti vartotojo vertybes ir pagal tai nustatyti įmonės prioritetus;
- pagal prioritetus supaprastinti ir racionalizuoti pagrindinius įmonės produktus ir paslaugas, kad skaitmenizuojant būtų atsisakyta esamo nereikalingo kompleksiškumo;
- susieti kuriamą vertę klientui su pagrindinėmis atliekamomis operacijomis ir sukurti naują operacinį modelį remiantis sudarytomis sąsajomis. Dirbama pirmiausiai analizuojant poreikius iš kliento pusės, tuomet naudojant skaitmeninius įrankius ir supaprastinant procesus;
- viena po kitos tobulinamos svarbiausios vartotojo kelionės keičiant operacinius procesus, kad būtų pagerintas efektyvumas ir greitis;

- naudojami lankstūs, įvairių funkcijų darbo metodai, pakeičiama valdymo sistema taip, kad būtų palaikomas nuolatinis tobulėjimas.

Apibendrinant, mokslinėje literatūroje yra įvairių vartotojo patirties apibrėžimų, tačiau šiame darbe vartotojo patirtis apibrėžiama kaip vidinis ir subjektyvus vartotojo atsakas į tiesioginį ir netiesioginį kontaktą su kompanija. Technologijos ir įmonių veiklos skaitmenizavimas keičia vartotojų patirtį ir kelia aukštesnius reikalavimus kompanijoms atitikti aukštus reikalavimus paprastumui, personalizavimui ir interaktyvumui. Geriausiai vartotojo patirtis pažįstama naudojant statinį ir dinaminį patirties suvokimą, o dinaminę patirtį realiu laiku stebėti ir matuoti geriausiai leidžia vartotojų stebėsena internete. Įvairūs teksto analizės metodai leidžia iš vartotojų atsiliepimų gauti naudingos informacijos apie vartotojo patirtį, o semantinė analizė atskleidžia esamus ir padeda prognozuoti būsimus vartotojo patirties aspektus.

1.5.Sentimentų analizės metodai

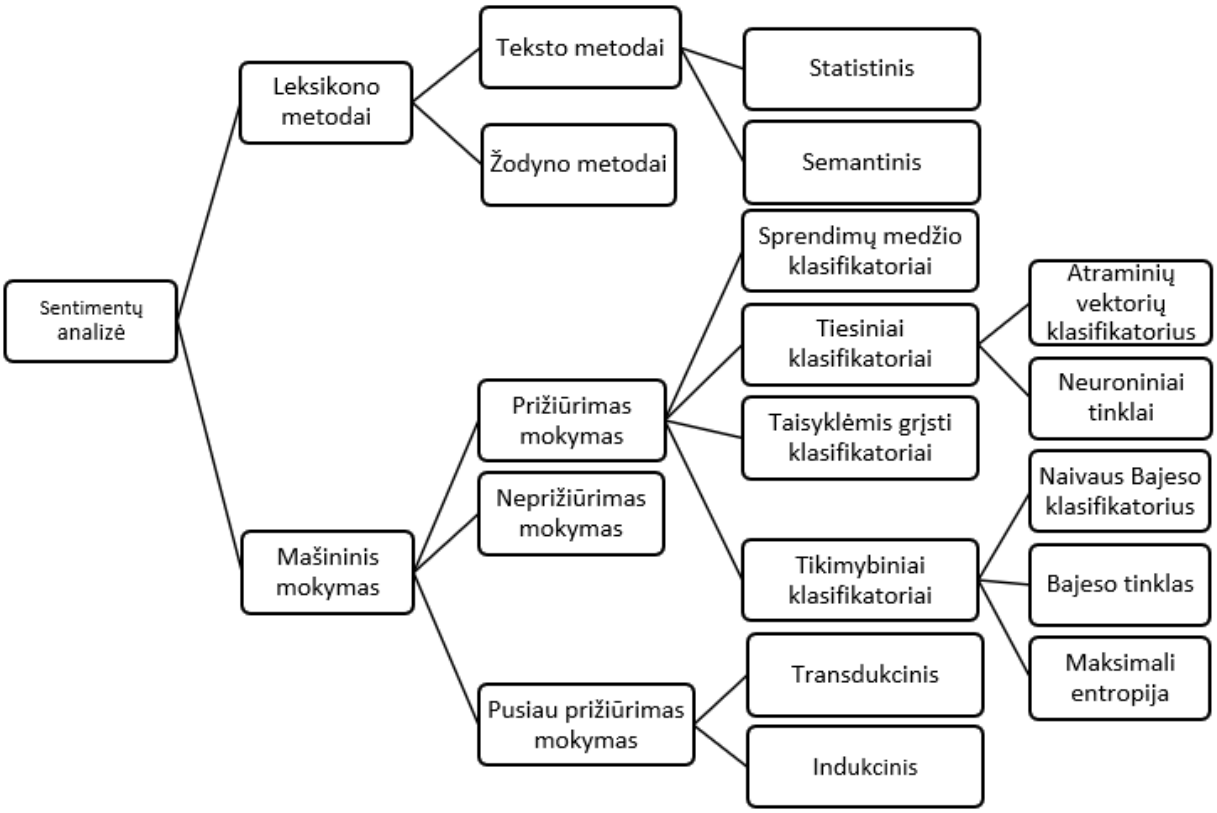
Sentimentų analizės uždaviniai skiriasi priklausomai nuo pritaikymo srities, toje srityje prieinamų žodynų kiekio ir kokybės, turimų kompiuterinių resursų, tiriamų tekstinių duomenų apimties – visi šie kriterijai lemia konkrečių teksto analizės metodų pasirinkimą. Šiame skyriuje aptariami egzistuojantys metodai, jų trūkumai ir privalumai. Sentimentų analizėje naudojami dažnai naudojami terminai paaiškinti 3 lentelėje [15].

3 lentelė. Sentimentų analizės terminai ir apibrėžimai

Terminas	**Apibrėžimas**
Žodžių krepšelio metodas (angl. *Bag of Words, BoW*)	Požymių išskyrimo metodas, kai žodžio panaudojimas tekste/atsiliepime panaudojamas kaip požymis nepaisant žodžių tvarkos ir gramatikos taisyklių.
Klasifikavimas	Mašininio mokymo procedūra, kuri padeda nustatyti grupę, kuriai priklauso naujas stebinys. Klasifikatorius – modelis, kuris turi būti apmokomas su mokymo duomenimis, kuriuose yra sužymėtos klasės. Apmokytas klasifikatorius gali klasifikuoti naujus stebinius.
Požymių išskyrimas (angl. *feature extraction*)	Procesas, kurio metu sudaromos arba atrenkamos reikšmingos, informatyvios ir išskiriančios vertės iš duomenų aibės, kurios panaudojamos klasifikatoriaus apmokyme.
N-grama	Žodžių arba simbolių junginys, sudarytas iš n gretimų žodžių arba simbolių, naudojamas modeliuoti žodžių / simbolių eiliškumą.
Unigrama	Specialus n-gramos atvejis, kai n=1.
Kalbos dalis (angl. *Part of speech, POS*)	Žodžių grupė, turinti panašias gramatines (sintaksines, morfologines) savybes, pavyzdžiui, daiktavardžiai, veiksmažodžiai, būdvardžiai, prieveiksmiai, jungtukai ir t.t.
Poliariškumas	Sentimento savybė turėti teigiamą arba neigiamą emociją.
Termino dažnis (angl. *Term frequency*, TF)	Termino dažnis – skaičius, parodantis, kiek kartų raidė arba žodis paminėtas tekste.

Termino dažnis – atvirkštinis dokumento dažnis (TF-IDF)	Termino dažnio ir atvirkštinio dokumento dažnio sandauga. Atvirkštinis dokumento dažnis parodo, ar terminas dažnai pasitiko visuose atsiliepimuose.
Silpnas sužymėjimas (angl. *weakly labelling*)	Duomenys, kuriems klasė buvo priskirta naudojant tam tikrą algoritmą, o ne žmogaus priskyrimą.

Mokslininkas F. H. Khanas savo darbe suskirstė sentimentų analizės metodus pagal tai, ar jie grįsti mašininiu mokymu, ar leksikonu (1 pav.). Leksikonu grįstiems metodams taikyti reikia iš anksto turėti žodyną arba jį susidaryti iš turimo tekstyno, tačiau tokiu atveju metodo tikslumas sumažėja [42]. Leksikonu grįsti metodai dažnai atsiliepimo poliariškumą apskaičiuoja pagal semantinę žodžių ar sakinių orientaciją. Semantinė orientacija – subjektyvumo tekste matas [43]. Iš mašininio mokymo algoritmų čia paminėti dažnai naudojami, tačiau sentimentų analizėje jų taikoma daugiau. Pastaraisiais metais itin dažnai taikomas neuroninių tinklų metodas [44].



2 pav. Sentimentų analizės metodai (sudaryta pagal Sun, Luo, ir Chen [44])

Taip pat metodai gali būti skirstomi pagal tai, ar matuoja sentimento stiprumą skirtingiems produkto aspektams, ar globaliai visam tekstui. Dauguma spendimų, besiremiančių globaliu vertinimu, įvertina tik teksto poliariškumą (teigiamas / neutralus / neigiamas) ir naudoja mašininio mokymo algoritmus. Sprendimai, orientuoti į detalesnį atsiliepimų klasifikavimą remiasi lingvistiniais aspektais tokiais kaip intensyvumas, neigimas, modalumas, kalbos dalys [43].

Ribeiro et. al visus sentimentų analizės metodus skiria į leksikono, mašininio mokymo ir hibridinius metodus [45]. Žemiau pateikiamos leksikono, mašininio prižiūrimo ir mašininio neprižiūrimo mokymosi, semantinio ir hibridinio metodų grupės.

Mašininio mokymosi metodai taikomi tam tikra žingsnių seka [15]:

1. duomenų surinkimas iš turimų šaltinių, pavyzdžiui, socialinių tinklų, tinklaraščių ir t.t.;
2. apdorojimas (angl. *pre-processing*) – duomenų valymas, segmentavimas, teksto padalijimas į teksto vienetus – leksemas (angl. *tokenization*) ir t.t.;
3. reikšmingų požymių išskyrimas;
4. klasifikatoriaus mokymas.

Pagrindiniai mašininio mokymo trūkumai: priklausomybė – metodai yra priklausomi nuo didelio duomenų kiekio ir nagrinėjamos srities; nenuoseklumas – taikant skirtingus metodus rezultatai gali labai skirtis, net jei tai nedideli patobulinimai; neskaidrumas – rezultatų argumentavimo procesas sunkiai įvykdomas, nes algoritmai grąžina tik rezultatą [46].

Prižiūrimas mašininis mokymas

Prižiūrimo mašininio mokymo metodai mokymuisi naudoja duomenis, kuriems žmogus yra priskyręs tam tikrą klasę arba klasė buvo priskirta naudojant tam tikrą algoritmą [15]. Atramos vektorių (angl. *Support vector machine, SVM*) ir Naivaus Bajeso (angl. *Naïve Bayes, NB*) metodai yra dažniausiai mokslinėje literatūroje naudojami mašininio mokymo metodai sentimentų analizei. Literatūroje teigiama, kad SVM metodas geriau klasifikuoja lyginant su kitais algoritmais, kai turima mokymo duomenų imtis yra nedidelė [47][48][49]. SVM klasifikatorius naudoja mokymosi duomenis rasti optimaliai hiperplokštumai, kuri kuo tiksliau atskirtų duomenis į atskiras klases. Naivaus Bajeso klasifikatorius yra tikimybinis metodas, kuris naudoja Bajeso teoremą klasifikatoriaus sprendimo taisyklėse ir prielaidą, kad naudojami požymiai yra nepriklausomi. Šiems metodams reikia mažiau duomenų nei neuroniniams tinklams.

Neprižiūrimas mašininis mokymas

Šie metodai nuo prižiūrimo mokymo skiriasi tuo, kad mokintis nereikia duomenų su iš anksto priskirtomis duomenų klasėmis. Dažniausiai tai būna klasterizavimo metodai, naudojami sentimentų analizėje, duomenų tyryboje, vaizdų analizėje. Klasterizavimo užduotis – sugrupuoti duomenis taip, kad jie klasterių viduje būtų kuo panašesni, o klasteriai tarpusavyje kuo labiau skirtųsi. Klasterizavimo metodai buvo naudojami trumpų tekstų analizėje [50][51]. C. Lin ir Y. He pasiūlė metodą, kuris naudoja latentinį Dirichlė paskirstymą (angl. *Latent Dirichlet Allocation, LDA*) sentimento ir temos išskyrimui [52]. Šis metodas pasiekia artimą prižiūrimo mokymosi metodams tikslumą.

Leksikonu grįsti metodai

Leksikonu arba žodynu grįsti metodai remiasi sudarytais sentimentų žodynais ir nustatytomis taisyklėmis. Sentimentų žodynai sudaromi žmonių, kompiuterių arba pusiau automatiškai. Sudarant žodynus sentimentai priskiriami žodžiams visiškai nevertinant konteksto [15]. Dažnai tokie žodynai sudaromi pradedant nedideliu skaičiumi žodžių su iš anksto žinomu sentimentu ir naudojantis didelės apimties žodynais surenkami sinonimai. Taip pat pasinaudojama natūralios kalbos apdorojimo (NLP) metodais, pavyzdžiui, lemavimu ar kalbos dalių išskyrimu [53].

H. Saif‘as et al. pasiūlė leksikonu grįstą semantinį metodą, vadinamą *SentiCircles*, kuris geba atnaujinti ir koreguoti žodžių sentimentų orientaciją remiantis jų pasikartojimo šablonais [54]. Kiti darbai įgalina NLP išskirti lingvistines savybes, tokias kaip sustiprinimas, neigimas, modalumas ir leksikonus identifikuoti dokumento [55] arba bendrą sentimentą [56] naudojant sentimento įverčio vidurkį.

Hibridiniai metodai

Hibridiniuose metoduose leksikonu ir mašininiu mokymu grįsti požiūriai, kurie gali gali lygiagrečiai vertinti sentimento poliariškumą. Gauti rezultatai agreguojami gaunant galutinį sentimento poliariškumą. Taip pat yra atlikta tyrimų, kai leksikonu ir mašininiu mokymu grįsti metodai taikomi skirtinguose modelio etapuose [57][58]. Hibridiniai metodai gali būti gauti derinant ir kitus algoritmus, pavyzdžiui, A. Tripathy, A. Anand‘as, and S. K. Rath‘as ištyrė, kad atliekant reikšmingų požymių atrinkimą su SVM metodu ir gautą rezultatą pateikiant dirbtiniam neuroniniam tinklui (ANN) kaip įvesties duomenis gaunamas aukštesnis tikslumas nei lyginant pavienių metodų rezultatus su tomis pačiomis filmų apžvalgomis iš IMDB duomenų bazės [59].

1.6.Giluminio mokymosi metodai

Sentimentų klasifikavimas dažnai laikomas atskiru dokumentų klasifikavimo atveju. Tokiu atveju dokumento konvertavimas į vektorių turi didelę reikšmę, nes privalo tiksliai atspindėti originalią informaciją pateiktą dokumente žodžiais arba sakiniais. Klasikinis teksto tyrybos metodas atvaizduoti tekstą į vektorių yra žodžių rinkinys, kuris laiko dokumentą kaip savo paties žodžių rinkinį. Remiantis šiuo rinkiniu dokumentas paverčiamas į fiksuoto ilgio skaitinį požymių vektorių, kurio elementai binariniai priklausomai, ar žodis tekste egzistuoja, dažnis arba TF-IDF įvertis. Vektoriaus dimensija lygi žodyno dydžiui. Paprastai BoW dokumento vektorius būna labai retas (kada daug reikšmių yra nuliai), nes vienas dokumentas turi tik mažą dalį viso žodyno žodžių. Pirmieji neuroniniai tinklai taip sudarydavo požymius [60].

BoW turi trūkumų, tokių kaip neatkreiptas dėmesys į žodžių tvarką, todėl du dokumentai gali turėti tą patį požymių vektorių, jei juos sudaro tie patys žodžiai, nors ir sudėti visai kitu eiliškumu. N-gramų rinkiniai yra BoW patobulinimas, kuris bando įvertinti žodžių tvarką nedideliame kontekste,

tačiau duomenų retumo ir didelio dimensijų skaičiaus problemos išlieka [60]. Taip pat, BoW neįvertina semantinės žodžių prasmės, pavyzdžiui, žodžiai „gražus“, „dailus“ ir „stalas“ pagal šį metodą turi tokį patį atstumą nuo vienas kito, tačiau semantiškai žodžiai „gražus“ ir „dailus“ turėtų būti arčiau nei „stalas“.

Šias problemas išspręsti buvo pasiūlyta taikyti žodžių vektorizavimo metodus (angl. *word embedding*), paremtus neuroniniais tinklais ir taip sukurti suspaustą vektoriaus reprezentaciją (angl. *dense vector*) arba mažo dimensijų skaičiaus vektorius, kurie gali atvaizduoti žodžius, jų semantines ir sintaksines savybes. Žodžių vektorizavimo technika naudojama kalbos modeliavimui ir požymių išgavimui, kada žodyno žodžiai transformuojami į realių skaičių vektorius. Daugelis gilių neuroninių tinklų tokius vektorius naudoja kaip įvesties požymių vektorius [61]. Tankūs vektoriai taip pat gali būti gaunami panaudojant žodžių vektorizavimo metodus jau sudarytam BoW.

E. Cambria, S. Poria, D. Hazarika ir K. Kwok‘as pasiūlė tekstą analizuoti principu viršus-apačia turint omenyje, kad semantini prasmei suprasti pirmiausiai reikia sakinių ir žodžių junginių prasmę ir tik tada leistis iki pavienio žodžio lygmens. Autorių pasiūlytas metodas koncepcinius bazinius elementus iš teksto ir susieja juos su bendrai žinomomis koncepcijomis naujoje trijų lygių sentimentų analizės žinių reprezentavimo sistemoje. Žodžiams ir jų junginiams pakeisti baziniais elementais taikomi rekurentiniai neuroniniai tinklai, tuomet daugiamačių skalių algoritmas (angl. *multi-dimensional scaling*) iš bazinių elementų sudaro bendrai žinomas koncepcijas [46].

W. Zhao et al. sukūrė naują gilaus neuroninio tinklo sistemą produktų atsiliepimų sentimentų klasifikavimui, kuris naudoja dažniausiai prieinamą reitingą kaip silpną duomenų sužymėjimą. Sistema sudaryta iš dviejų dalių: (1) mokoma atpažinti bendrą sentimentų pasiskirstymą sakiniuose naudojant reitingo informaciją; (2) ant paslėptųjų sluoksnių pridedamas klasifikavimo lygmuo ir tuomet turimi žmogaus sužymėti atsiliepimai naudojami papildomam prižiūrimam mokymuisi. Pirmajam lygmeniui naudojami konvoliuciniai neuroniniais tinklai požymių išgavimui ir LSTM (angl. *Long Short-Term Memory)* algoritmas [62].

Y. Ma, H. Peng‘as ir E. Cambria taip pat pasiūlė patobulintą LSTM metodo versiją Sentic LSTM, naudojančią bendrai visuomenėje žinomas koncepcijas sentimentų analizei aspekto lygmeniu. Metodas patobulina LSTM hierarchiniu mechanizmu, susidedančiu iš objekto lygio ir sakinio lygio analizės. Bendrai visuomenėje žinomos koncepcijos naudojamos gilaus neuroninio tinklo apmokymui [63].

Young‘as et al. pasiūlė kitą hierarchinio dėmesio sistemą dokumento lygio sentimentams vartotojų atsiliepimuose prognozuoti. Modelį sudaro dviejų lygių dėmesio mechanizmai: (1) žodžio lygmuo, (2) sakinio lygmuo, kurie leidžia modeliui, priklausomai nuo esamos situacijos, daugiau dėmesio skirti žodžiams arba sakiniams sudarant dokumento atvaizdavimą [64]

Xu et al. pasiūlė trumpalaikės atminties (angl. *cashed*) LSTM modelį, kuris skirtas rasti bendrą sentimentą didelės apimties tekstiniame dokumente. Šiame metode atmintis padalijama šį keletą grupių su skirtingais pamiršimo koeficientais. Metodas įgalina grupes su nedideliu pamiršimo koeficientu surasti globalaus sentimento požymius, o grupes su nedideliu pamiršimo koeficientu rasti lokalaus sentimento požymius [65].

Y. Zhang‘as et al. pasiūlė prieštaringos atminties tinklą skirtingų sričių ir tematikų sentimentų klasifikavimui. Autorių siūlomas metodas geba perkelti mokymosi su šaltinio duomenimis rezultatą į tikslinę sritį. Jis paraleliai apmoko du klasifikatorius: sentimentų klasifikavimui ir srities klasifikavimui, t.y. atskirti, ar dokumentas yra iš šaltinio ar tikslinės srities. [66]

Pastaraisiais metais išaugusi gilaus mokymosi tyrimų ir taikymų skaičiui sentimentų analizės srityje yra svarbu pagal poreikius atsirinkti tinkamą metodą, suprasti galimų alternatyvų privalumus ir trūkumus bei žinoti, kurie metodai geriausiai veikia tam tikram uždaviniui. Todėl tolimesnėje analizėje pasirinkta palyginti naujus gilaus mokymosi metodus naudojant įvairių sričių ir apimčių duomenų rinkinius.

2. Tyrimų metodai

Šioje darbo dalyje pateikiama tyrime naudojamų metodų teorinė medžiaga. Pirmiausiai aprašomi duomenų vektorizavimo metodai: BoW, TF-IDF, Doc2Vec_dbow, Doc2Vec_dbow_w, Doc2Vec_dm_m, Sent2Vec (fastText), LSI, RP. Taikomi trys mašininio mokymo algoritmai: logistinė regresija, atsitiktiniai miškai ir neuroniniai tinklai ir 5 žodynais grįsti metodai: Harvard-IV žodynas, Henry's finansų srities žodynas (Journal of Business Communication), Loughran-McDonald finansų srities žodynas, QDAP Vartotojo sentimento žodynas, SenticNet. Modelių efektyvumui įvertinti ir palyginti naudojama sumaišymo matrica ir kitos apskaičiuotos metrikos.

2.1.Požymių išskyrimo algoritmai

Siekiant spręsti sentimento detekcijos uždavinį didelę rezultato dalį lemia sėkmingas klasifikavimo požymių sudarymas. Klasifikavimui naudingi požymiai gali būti sudaromi remiantis žodžių krepšelio metodu, ontologijomis ir modeliais. Mašininio mokymosi algoritmai požymiais laiko iš žodžių, sakinių arba dokumentų sudarytus vektorius. Žodyno metodo atveju žodžiai yra įvertinami kiekybiniu sentimento klasės, kuriai priklauso, įverčiu.

Reikšmingų klasifikavimo požymių atranka yra sudėtingas uždavinys, nes reikia sumažinti dimensijų skaičių taip, kad būtų pasiektas geriausias klasifikavimo rezultatas atsižvelgiant į klasifikavimo tikslumą ir skaičiavimo resursų optimizavimą. Šiame skyriuje aptariami tyrime naudoti reikšmingų požymių sudarymo ir atrankos metodai.

2.1.1. Žodžių rinkinio krepšelis

Atliekant teksto analizę pirmiausiai tekstas turi būti paverčiamas į vektorių – šiai užduočiai dažniausiai naudojamas žodžių rinkinio krepšelio metodas BoW. Tai yra toks požymių išskyrimo metodas, kai žodžio vartojimas tekste/atsiliepime panaudojamas kaip požymis nepaisant žodžių tvarkos ir gramatikos taisyklių. Šiuo požiūriu kiekvieno žodžio pavartojimo dažnis yra kaip požymis. BoW gali būti labai paprastas arba sudėtingas taikymas priklausomai nuo to, kaip sudaromas žodynas žinomų žodžių arba leksemų žodynas ir kaip įvertinamas žinomo žodžio dokumente.

Žodžių krepšelis sudaromas iš turimų tekstinių duomenų išrenkant unikalius žodžius – sudaromas žodynas ir įvertinamas jo žodžių pasikartojimo dažnis kiekviename dokumente. Tikslas – iš dokumento sukurti vektorių, kurį būtų galima panaudoti kaip įvesties duomenis mašininio mokymo algoritmui. Kadangi skirtingų žodžių dažnai būna daugiau nei nagrinėjamų dokumentų, siekiant išvengti daugybės dimensijų, pasirenkamas fiksuotas dimensijų skaičius. Dokumentai paverčiami dvinariais vektoriais su 1 reikšme, jei žinomas žodis yra dokumente ir 0, jei nėra. Nauji dokumentai,

kuriuose pasitaiko žodyne nesančių žodžių, vis tiek gali būti paverčiami vektoriais ignoruojant nežinomus žodžius.

Siekiant išvengti daugybės dimensijų prieš reprezentuojant dokumentus vektoriais gali būti atliekama teksto transformacija: visos raidės paverčiami mažosiomis, pašalinami skyrybos ženklai, sudaromos leksemos – teksto vienetai, ignoruojami dažnai pasitaikantys ir daug prasmės nesaugantys žodžiai, pavyzdžiui, dalelytės, sutvarkomi klaidingai parašyti žodžiai, atliekamas lemavimas. Sudėtingesnis būdas yra į žodyną įtraukti ir n-gramas.

Vietoj dvinarės reikšmės į vektorių gali būti įrašomas pasitaikančių žodžių dažnis arba šio dažnio santykis su visais dokumente esančiais žodžiais.

2.1.2. Termino dažnis – atvirkštinis dokumento dažnis

Termino dažnis – atvirkštinis dokumento dažnis (TF-IDF) – metrika, kuri parodo žodžio svarbą dokumente viso dokumentų rinkinio kontekste. Ši metrika dažnai naudojama kaip svorinis koeficientas ir yra tiesiogiai proporcinga žodžio panaudojimo dažniui dokumente ir atvirkščiai proporcinga žodžio panaudojimo dažniui dokumentų rinkinyje.

TF-IDF yra termino dažnio dokumente sandauga su atvirkštiniu dokumento dažniu. Terminu gali būti laikomas žodis, jo bendrinė forma, n-grama arba kitas pasirinktas teksto vienetas. Abiem statistikoms apskaičiuoti yra keletas būdų, tačiau paprasčiausia yra apskaičiuoti, kaip dažnai terminas t_i pasitaiko dokumente arba dokumentų rinkinyje. Jei terminas pasitaiko n_i dokumentuose, tai TF-IDF apskaičiuojama pagal (1) formulę [67]:

$$idf(t_i) = log \frac{N}{n_i}. \quad (3)$$

Aukštas TF-IDF rodiklis reiškia, kad terminas turi didelį dažnį dokumente ir mažą panaudojimo dažnį dokumentų rinkinyje, todėl pagal šio rodiklio svorį dažnai išfiltruojami bendriniai, dažnai pasitaikantys žodžiai.

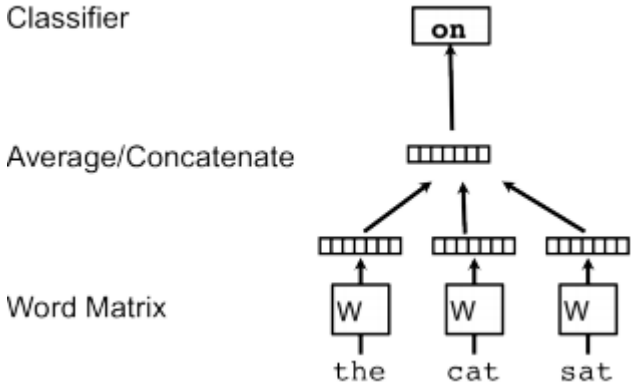
2.1.3. Žodžių vektoriai

Paprasčiausias žodžio vektorius – „one-hot“ vektorius, kuris išreiškia kiekvieną žodį kaip $R^{|V|\times 1}$ vektorius su visomis komponentėmis lygiomis nuliui ir vienu vienetu tame požymyje, kuriam priskirtas vektorizuojamas žodis. Žodžių vektoriai šiuo būdu atrodytų taip [68]:

$$w^{aardvark} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, w^{a} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, w^{at} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \cdots w^{zebra} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \quad (4)$$

Kiekvienas žodžio vektorius yra nepriklausomas nuo kitų vektorių, todėl šis metodas neduoda tiesioginio panašumo tarp žodžių įverčio. Dėl šios priežasties jau sudarytiems „one-hot“ vektoriams taikomi kiti metodai vektorių erdvei sumažinti ir tarpusavio panašumams rasti.

Žodžių vektorių apmokymo metode kiekvienas žodis sudaro unikalų vektorių-stulpelį matricoje W. Stulpelio indeksas priklauso nuo žodžio vietos žodyne. Vektorių suma panaudojama kaip požymis prognozuoti šalia esantį žodį sakinyje. Žodžių vektorių apmokymo sistema pavaizduota 1 paveiksle.



3 pav. Žodžių vektorių apmokymo metodas
Šaltinis: T. Mikolov‘as [69]

Taigi, turint mokymosi žodžių seką w_1, w_2, w_3, ..., w_T žodžių vektorių modelio tikslas – maksimizuoti vidutinę tikimybę:

$$\frac{1}{T}\sum_{t=k}^{T-k} \log p(w_t|w_{t-k}, ..., w_{t+k}) \quad (5)$$

Prognozė dažniausiai atliekama naudojant daugialypį klasifikatorių, pavyzdžiui, naudojant paprastą arba hierarchinę minkštojo maksimumo (angl. *softmax*) aktyvacijos funkciją. Google sukurtas neuroniniais tinklais grįstas žodžių vektorių sudarymo algoritmas *word2vec* naudoja stochastinį gradientinio nusileidimo metodą, kai gradientas apskaičiuojamas naudojant atgalinio skleidimo (angl. *backpropagation*) algoritmą.

2.1.4. fastText metodo modifikacija – Sent2Vec

Algoritmas *fastText* yra viešai prieinamas *Facebook* dirbtinio intelekto tyrėjų sukurtas įrankis teksto pavertimui vektoriais ir klasifikavimui. Šis algoritmas skiriasi nuo anksčiau apžvelgtų tuo, kad mažiausia teksto dalimi, kurią galima paversti vektoriaus požymiu yra laikomas ne žodis, o pasirinkto

dydžio žodžio raidžių n-grama. Pavyzdžiui, žodžio „lapas“ 3-grama būtų: „lap“, „apa“, „pas“. Tuomet žodžio „lapas“ vektorius būtų visų n-gramų kombinacija. Tokio algoritmo privalumai (pagal [70]):

1. retų žodžių pavertimas vektoriais;
2. žodžių, nesančių pradiniame žodyne, vektorizavimas;
3. *fastText* vektorizuoja mažus duomenų rinkinius geriau nei *word2vec*.

Modelis:

Kiekviena žodis w yra laikomas raidžių n-grama. Jei turime G dydžio n-gramų žodyną, tai žodyje w esantis n-gramų rinkinys yra $G_w \subset \{1, \dots, G\}$. Kiekvienai n-gramai g priskiriama vektorinė išraiška z_g. Žodis išreiškiamas kaip visų jo n-gramų vektorių suma:

$$s(w, c) = \sum_{g \in G_w} z_g^T v_c \tag{5}$$

čia v – konteksto žodžio vektorius

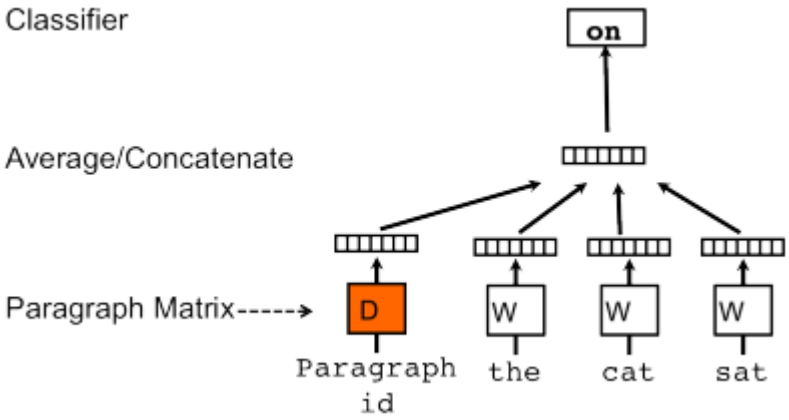
c – konteksto žodžio indeksas konteksto žodžių rinkinyje.

Siekiant apriboti atminties dydį n-gramoms priskiriami sveikieji skaičiai nuo 1 iki K. Galiausiai žodis išreiškiamas indeksu žodyne ir jį sudarančių n-gramų aibe [70].

2.1.5. Pastraipų vektorius – paskirstytos atminties modelis

Pastraipų vektorius – paskirstytos atminties modelis PV-DM (angl. *paragraph vector – distributed memory*) yra patobulinta žodžių vektorių apmokymo versija.

Pastraipų vektoriaus modelis gali konvertuoti į vektorių bet kokio ilgio tekstus: sakinius, pastraipas ir dokumentus. Jis nekelia reikalavimo, kad nagrinėjamai sričiai būtų specialiai apmokinta svorių funkcija. Šiame metode kiekviena pastraipa yra atvaizduojama unikaliu vektoriumi-stulpeliu matricoje D ir kiekvienas žodis taip pat atvaizduojamas unikaliu vektoriumi matricoje W. Pastraipos ir žodžių vektoriams paskaičiuojamas vidurkis arba jie suliejami (angl. *concatenated*) ir prognozuojamas sakinyje šalia esantis žodis. PV-DM pastraipos vektorių apmokymo sistema pavaizduota 2 paveiksle [69].



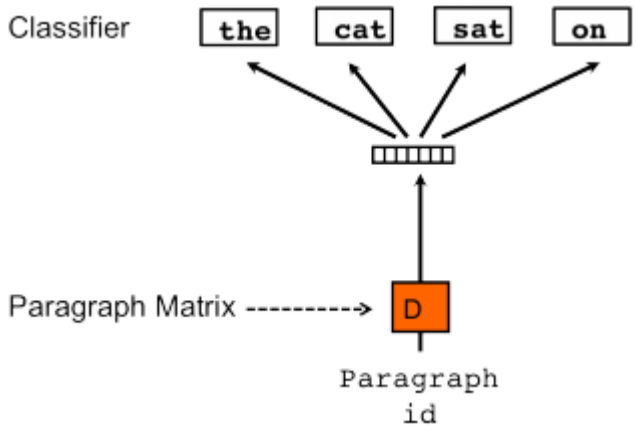
4 pav. PV-DM pastraipos vektorių modelio apmokymas: Le ir Mikolov [69]

Pastraipos vektorius gali būti laikomas tokiu pačiu kaip ir žodžio vektorius, tačiau jis saugo informaciją apie trūkstamą konteksto dalį, t.y. informaciją apie pastraipos temą. Kontekstu vadinamas pasirinkto fiksuoto dydžio žodžių rinkinys. Pastraipos vektorius yra vienodas visiems tos pastraipos konteksto rinkiniams, tačiau skiriasi kitose pastraipose. Žodžių vektorių matrica W yra ta pati visoms pastraipoms. Pastraipos ir žodžių vektoriai kaip ir prieš tai buvusiame algoritme, apmokomi naudojant gradientinį nusileidimo metodą [69].

Pagrindinis pastraipos vektorių metodo pranašumas yra tai, kad jie yra apmokomi naudojant duomenis be iš anksto priskirtų žymių (angl. *labels*), todėl gali puikiai susitvarkyti su užduotimis, kuriose yra mažas sužymėtų duomenų kiekis. Taip pat metodas leidžia išsaugoti žodžių semantiką, pavyzdžiui, vektorių erdvėje angliškas žodis „powerful“ yra artimesnis žodžiui „strong“ nei „Paris“. Be to, šis pastraipos vektorių metodas įvertina žodžių tvarką bent jau nedideliame kontekste ir neturi tokios problemos kaip n-gramų modelis, kuris taip pat įvertina žodžių tvarką, tačiau dėl to sukuria labai daug dimensijų turintį vektorių ir negali gauti gerų rezultatų skirtingoms sritims [69].

2.1.6. Pastraipų vektorius – paskirstyto žodžių krepšelio metodas

Pastraipų vektorius – paskirstyto žodžių krepšelio metodas PV-DBOW (angl. *Paragraph Vector – Distributed Bag of Words*) yra paprastesnė praeitame skyriuje aptarto metodo versija, kurioje ignoruojami visi konteksto rinkinyje esantys žodžių vektoriai ir žodžiai prognozuojami iš pastraipos atsitiktiniu būdu. Taigi, kiekvienoje stochastinio gradientinio nusileidimo metodo iteracijoje iš pastraipos paimama pasirinkto dydžio žodžių rinkinys ir iš jo atsitiktiniu būdu išrenkamas žodis, kuriam pagal pastraipos vektorių prognozuojami šalia esantys žodžiai. PV-DM pastraipos vektorių apmokymo sistema pavaizduota 3 paveiksle [69].



5 pav. PV-DBOW pastraipos vektorių modelio apmokymas: Le ir Mikolov [69]

Šis metodas naudoja nesudėtingą logiką ir saugo mažesnį duomenų kiekį, nes reikia saugoti tik minkštojo maksimumo aktyvacijos funkcijos parametrus ir nereikia saugoti žodžių vektorių matricos.

2.1.7. Latentinė semantinė analizė

Latentinė semantinė analizė (LSA), dar žinoma kaip latentinis semantinis indeksavimas (LSI), yra metodas nustatyti ryšiams tarp dokumentų ir juose esančių sąvokų bei sudaryti koncepcijų matricą, siejančią dokumentus ir sąvokas. Metodas remiasi TD-IDF ir ypatingųjų reikšmių dekompozicija (angl. *singular-value decomposition,* SVD). SVD metodas ne tik padaro suspaustą matricos reprezentaciją, bet ir pašalina originalioje sąvokų-dokumentų matricoje pasitaikantį triukšmą. SVD metodas plačiau žinomas dėl pritaikymo principinių komponenčių analizėje.

Pirmiausiai sudaroma matrica visiems dokumentams (atvaizduojama stulpeliais) ir juose esantiems žodžiams (atvaizduojama eilutėmis), tuomet jos dimensijų skaičiui sumažinti ir gauti informatyvesnius vektorius taikoma SVD, kuri išlaiko tą pačią ryšių tarp stulpelių struktūrą:

$$A = USV^T; \tag{6}$$

čia A – n×p įėjimo duomenų matrica (LSI atveju matricos eilutę atitinka dokumentas, paverstas į TF-IDF vektorių); U – n×p ortogonali matrica; S – p×p diagonali matrica (angl. *singular values*); V – p×p ortogonali matrica (čia eilutė atitinka principinės komponentės koeficientų vektorių).

Tuomet galima palyginti žodžių panašumą apskaičiuojant kampo kosinusą tarp dviejų eilučių vektorių arba sandaugą tarp normalizuotų vektorių reikšmės. Rezultatas, artimas vienetui, reiškia labai panašius dokumentus, o artimas nuliui – labai nepanašius dokumentus.

2.2. Mašininiu mokymu grįsti klasifikavimo metodai

2.2.1. Logistinė regresija

Dvilypės logistinės regresijos metodu galima ištirti binarinio kintamojo priklausomybę nuo bet kokio tipo aiškinamųjų kintamųjų (kategoriniai kintamieji transformuojami). Binarinės regresijos metodas leidžia įvertinti prognozuojamą kintamąjį tolydžių kintamųjų skalėje ir tai reiškia tikimybę, kad prognozuojamas įvykis įvyks. Tikimybė, kad prognozuojamas dydis Y_i įgis reikšmę 1, t.y įvykis įvyks, skaičiuojama šia formule :

$$p_i = \frac{e^{z(x_i)}}{1 + e^{z(x_i)}}, z(x_i) = a + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_n x_{ni} \quad (7)$$

Čia $x_{1i}, \dots, x_{ki}$ — nepriklausomų kintamųjų reikšmės.

Duomenys turi tenkinti tokius reikalavimus [71]:

1. Priklausomas kintamasis Y turi būti dvireikšmis. Visi kiti kintamieji gali būti arba intervaliniai, arba dvireikšmiai (įgyti reikšmes 0 arba 1).
2. Duomenyse neturi vyrauti viena iš Y reikšmių. Keliami reikalavimai, kad tarp Y reikšmių vienetų (nulių) būtų ne mažiau penktadalio.
3. Nepriklausomi kintamieji negali stipriai koreliuoti. Stipriai koreliuojantys regresoriai gali iškreipia modelio priklausomybes.

Tegul priklausomas kintamasis priklauso nuo n nepriklausomų kintamųjų

$$x = (x_1, x_2, \dots, x_n) \quad (8)$$

Logistinės regresijos modelis išreiškiamas:

$$p_i = \frac{e^{z(x_i)}}{1 + e^{z(x_i)}} \quad (9)$$

čia e = 2,718…, $z(x_i) = a + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_n x_{ni}$.

Į reiškinį $z(x_i) = a + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_n x_{ni}$ įstačius $x_1, x_2, \dots, x_k$ reikšmes galima prognozuoti Y. Jeigu $z(x_i) > 0$, prognozuojama, kad $Y = 1$, jeigu $z(x_i) < 0$, tai prognozuojama, kad $Y = 0$, jeigu $z(x_i) = 0$, rekomenduojama sprendimą priimti metant monetą.

2.2.2. Atsitiktinis miškas

Atsitiktinis miškas (angl. *random forest,* RF) – yra medžio tipo klasifikatorių kolektyvas (angl. *ensemble*), kuris veikia apmokydamas daugybę atskirų sprendimo medžių ant skirtingų duomenų imčių, paprastai gaunamų savirankos (angl. *boostraping*) būdu, ir daugumos balsavimo būdu išrinkdamas dažniausiai sprendimo medžių prognozavimo rezultatuose pasitaikiusią klasę. Šis

metodas neturi sprendimo medžiui būdingos persimokymo problemos, kai modelis pernelyg tiksliai veikia su mokymosi duomenimis, bet prastai prognozuoja jam pateikus nematytus duomenis.

Atsitiktinis miškas yra žinomiausias savirankos agregavimo variantas (angl. *bagging*). Jei $X = x_1, ..., x_n$ yra apmokymo duomenų aibė su prognozuojamu kintamuoju $Y = y_1, ..., y_n$, tai vidurkinimo algoritmas taiko duomenų imties ėmimo metodą su pasikartojimu, t.y. ties patys x_i stebiniai gali pakliūti atsitiktinai atrinktą duomenų imtį daugiau nei vieną kartą. Pasirinktam skaičiui B medžių apmokyti atsitiktinai atrenkamos duomenų imtys iš X ir modelis apmokomas tokiu algoritmu:

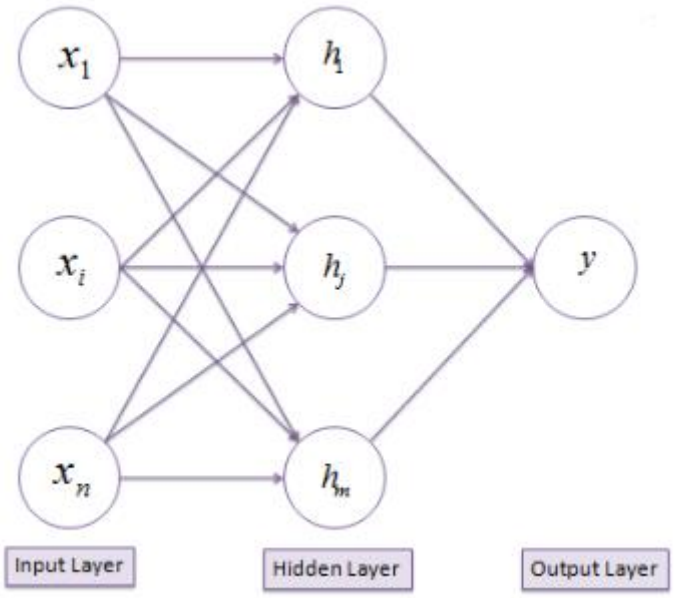
kiekvienam $b = 1, \dots, B$:

1. Atsitiktinis n dydžio duomenų imties X_b, Y_b ėmimas su pasikartojimu iš X, Y imties;
2. Su imtimis imties X_b, Y_b apmokomas sprendimų medis f_b.

Po apmokymo apskaičiuojama, kokia klasė buvo dažniausiai prognozuota ir tai yra bendras atsitiktinio miško metodo rezultatas. Dėl daugybės naudojamų sprendimo medžių rezultatų apibendrinimo šis metodas nėra jautrus triukšmui duomenyse, jei užtikrinami tarpusavyje nekoreliuoti sprendimo medžio modeliai. Tai užtikrinama apmokymui naudojant vis kitą pradinės apmokymo duomenų aibės imtį ir atsitiktinę modelio požymių imtį. Pagrindiniai metodo parametrai, kirie turi didžiausią įtaką meodelio rezultatui yra merdžių skaičius ir atsitiktinės modelio požymių imties dydis.

2.2.3. Neuroniniai tinklai

Daugiasluoksnis perceptronas (MLP) – prižiūrimojo mokymosi algoritmas, kuris apmokina funkciją $f: R^m \rightarrow R^o$ naudojant m dimensijų apmokymo imtį ir grąžinant o dimensijų rezultatą. Jei $X = x_1, ..., x_n$ yra apmokymo duomenų aibė ir Y yra prognozuojamas kintamasis, tai MLP gali apmokinti netiesinį klasifikatorių. Metodas nu logistinės regresijos skiriasi tuo, kad tarp įvesties ir išvesties sluoksnių gali būti vienas arba daugiau netiesinės funkcijos sluoksnių, vadinamų paslėptaisiais sluoksniais. paveiksle pavaizduotas vieno paslėptojo sluoksnio neuroninis tinklas [73].



6 pav. Vieno paslėptojo sluoksnio daugiasluoksnis perceptronas

Kairėje pusėje pavaizduotas sluoksnis vadinamas įvesties sluoksniu, sudarytas iš $\{x_i|x_1,\ x_2, \dots, x_n\}$ neuronų, reprezentuojančių įvesties duomenų aibės požymių skaičių. kiekvienas neuronas paslėptajame sluoksnyje transformuoja vertes iš praeito sluoksnio tiesine suma su svoriniais koeficientais $w_1x_1 + w_2x_2 + \dots + w_nx_n$ ir panaudoja netiesinę aktyvacijos funkciją. Išvesties sluoksnis gauna reikšmes iš paskutiniojo paslėpto sluoksnio ir transformuoja jas į galutinį rezultatą. Tyrime naudojamas MLP apsimokymui taiko atgalinio skleidimo metodą. Parametrų apmokymui naudojama kryžminės entropijos klaidos funkcija [73]:

$$Loss(\hat{y}, y, W) = -y \ln \hat{y} - (1 - y) \ln (1 - \hat{y}) + \alpha ||W||_2^2 \quad ; \qquad (10)$$

čia y – tikslinis kintamasis;

$\hat{y}$ - prognozuojamas kintamasis;

$\alpha||W||_2^2$ - L2 reguliarizacijos narys;

α > 0 – hiperparametras, keičiantis reguliarizacijos bausmės dydį.

MLP jautriai reaguoja į įvesties duomenų skalę, todėl siekiant tikslesnių rezultatų patariama transformuoti duomenis į [0, 1], [-1, +1] skalę arba standartizuoti duomenis.

2.3.Žodynu grįsti klasifikavimo metodai

Šiame skyrelyje apžvelgiami internete laisvai prieinami įrankiai sentimentų analizei atlikti, tokie kaip R programavimo aplinkos paketo *SentimentAnalysis* siūlomi parengti žodynai.

SenticNet

SenticNet programa yra laisvai prieinamas įrankis sentimentų analizei, kuriuo galima naudotis per nepriklausomą XML repozitoriją arba naudojant aplikacijų programavimo sąsają (API). SenticNet yra semantikos resursas koncepcinio lygio sentimentų analizei, kuris naudoja grafų teoriją ir daugiamačių skalių metodus sujungti žodžio lygmens natūralios kalbos duomenis su koncepcinio lygio nuomonėmis ir sentimentais. SenticNet žinių bazė gali būti naudojama įvairiose srityse, pavyzdžiui didžiųjų socialinių duomenų analizei, žmogaus ir kompiuterio sąveikai tobulinti, e-sveikatos srityje ir t.t. šis įrankis teikia sąsajas tarp 30 000 gerai žinomų koncepcijų, sudarytų iš pavienių žodžių arba n-gramų. Sentimentai skirstomi į klases: stipriai neigiami, silpnai neigiami, neutralūs, silpnai teigiami, stipriai teigimai. Skirtingai nuo kitų semantinės analizės įrankių SenticNet nėra sudarytas iš rankiniu būdu sužymėtų duomenų. Čia taikant grafus ir dimensijų mažinimo metodus surinkti duomenys pateikiami trimis būdais: semantiniu tinklu, matrica ir vektorių erdve. Semantinė ir sentimento reikšmės apskaičiuojamos derinant neuroninių tinklų ir emocijų kategorizavimo modelius [74].

SentimentAnalysis

R pakete SentimentAnalysis galima analizei naudoti specifinius žodynus, sudarytus specialiai tam tikrai sričiai. Žodynai sudaryti naudojant Bajeso metodus [75]. Tyrime naudojami žodynai:

- *Harvard-IV* žodynas (General Inquirer) – žodynas turi 1915 teigiamus ir 2291 neigiamus sentimentus išreiškiančius žodžius
- *Henry's* finansų srities žodynas (Journal of Business Communication)
- *Loughran-McDonald* finansų srities žodynas
- QDAP Vartotojo sentimento žodynas

2.4.Klasifikavimo kokybės įvertinimo metrikos

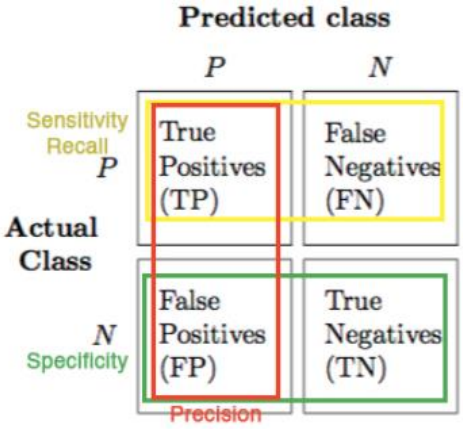
Apmokius modelį ir suklasifikavus duomenis reikia patikrinti, ar gautas rezultatas yra patikimas ir išsirinkti geriausią modelį. Šiuo tikslu galima sudaryti sumaišymo matricą ir skaičiuoti įvertinimo metrikas.

Pirmiausiai apsibrėžiame sąvokas, kas yra teigiama ir neigiama reikšmė. Teigiama reikšmė – tai tikslinė, labiausiai rūpinti reikšmė, pavyzdžiui, pardavėjui tai būtų pirkimas, medicininiuose tyrimuose tai būtų atsakymas, kad žmogus serga tam tikra liga ir t.t. neigiama reikšmė šiuo atveju būtų atitinkamai nepirks arba neserga. Šio tyrimo atveju tikslinė klasė – pozityvus sentimentas. Suklasifikavus testavimo duomenų rinkinio objektus gauname prognozes. T – teigiamai suklasifikuotos reikšmės, o N – neigiamai. Palyginame realias ir modelio apskaičiuotas reikšmes ir sudarome sumaišymo matricą, kurioje yra apskaičiuoti keturi dydžiai:

- teisingai teigiamas (TP) – teisingai suklasifikuotas teigiamas atvejas;

- teisingai neigiamas (TN) – neigiamos reikšmės, teisingai suklasterizuotos;
- neteisingai teigiamas (FT) – teigiamos reikšmės, neteisingai suklasterizuotos;
- neteisingai neigiamas (FN) – neigiamos reikšmės, neteisingai suklasterizuotos.

Sudaryta sumaišymo matrica (angl. *confusion matrix*) yra labai naudingas įrankis modelio vertinimui. Sumaišymo matricos tikslas yra identifikuoti, kokios rūšies klaidomis pasižymi apmokytas modelis, nes kai kurios rūšies modelių klaidos yra svarbesnės nei kitos tam tikru atveju.



7 pav. Sumaišymo matricos pavyzdys

Apibrėžiame kelias metrikas, pagal kurias taip pat naudinga įvertinti klasifikatorių.

Klasifikatoriaus tikslumas (angl. *accuracy*) apskaičiuojamas kaip procentinė dalis teisingai suklasifikuotų testavimo duomenų rinkinio objektų tarp viso testavimo rinkinio. Šis rodiklis parodo, kaip gerai bendrai klasifikuojami duomenys. Tikslumas efektyviausias tada, kai duomenys subalansuoti.

$$bendras\ tikslumas = \frac{TP\ sk. + TN\ sk.}{bendras\ objektų\ sk.} \tag{11}$$

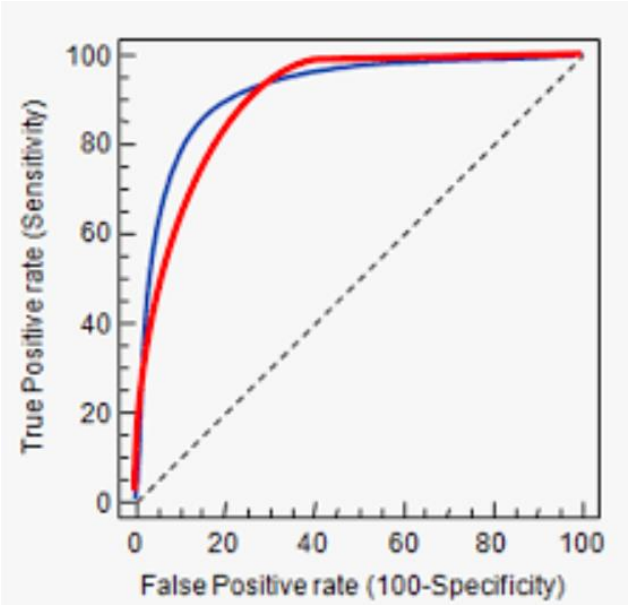
Kai klasės yra nesubalansuotos, labai maža dalis objektų priklauso teigiamai klasei, todėl vertinant tikslumą negalime įvertinti būtent tų kelių tikslinių objektų, kurių teisingu klasifikavimu esame labiausiai suinteresuoti. Tam padeda kitos metrikos, tokios kaip specifiškumas ir jautrumas. Specifiškumas parodo, kokia dalis neigiamų objektų suklasifikuota teisingai, o jautrumas – kokia dalis teigiamų.

$$specifiškumas = \frac{TN\ sk.}{TN\ sk. + FT\ sk.} \tag{11}$$

$$jautrumas = \frac{TP\ sk.}{TP\ sk. + FT\ sk.} \tag{12}$$

F1 įvertis yra tikslumo ir atkūrimo harmoninis vidurkis, skirtas apibendrinti abu matus. Kuo metrikos aukštesnė reikšmė, tuo geriau metodas klasifikuoja.

Taip pat naudojama daugybė kitų metrikų ir grafikų, pavyzdžiui, ROC kreivė. ROC kreivė (angl. *Receiver operating characteristic*) – grafikas, rodantis klasifikatoriaus jautrumo ir specifiškumo sąryšį (žr. 8 pav.). Atsitiktinai išvadą spėjančio klasifikatoriaus ROC kreivė yra įstrižai grafiką kertanti tiesė. Klasifikatorius tuo geresnis, kuo aukščiau šios tiesės yra jo kreivė. Jei kreivė atsiduria žemiau atsitiktinio klasifikatoriaus tiesės, tai reiškia jog toje srityje klasifikatorius klysta dažniau nei teisingai prognozuoja.



8 pav. ROC kreivės pavyzdys

Taip pat skaičiuojamas AUC (angl. *Area Under Curve*) – plotas po ROC kreive. Kuo metrikos aukštesnė reikšmė, tuo geriau metodas klasifikuoja. Tikimybinė AUC interpretacija: AUC yra tikimybė, kad atsitiktinai paimtiems teigiamos ir neigiamos klasės egzemplioriams, detektoriaus išėjimas teigiamos klasės atveju bus didesnis už detektoriaus išėjimą neigiamos klasės atveju.

3. Tyrimų rezultatai ir jų aptarimas

Šioje darbo dalyje pateikiami atliktų tyrimų rezultatai. Modeliai sudaryti naudojantis *Python* programavimo kalba ir jos paketais bei *Jupyter Notebook* programine įranga. Iš pradžių tekstiniai duomenys buvo paversti į vektorius, po to pritaikyti skirtingi klasifikavimo modeliai. Kiekvienas modelis įvertintas naudojantis sumaišymo matrica, įvairiomis metrikomis ir palygintos ROC kreivės. Palyginimui duomenys taip pat buvo klasifikuojami naudojant 5 žodynu grįstus modelius iš R paketo *SentimentAnalysis* ir sentimentų analizės programos *SenticNet*.

3.1.Tyrime naudojami duomenų rinkiniai ir taikomi modeliai

Siekiant sudaryti ir apmokyti modelius su skirtingais duomenų rinkiniais buvo pasirinkti 3 vartotojų atsiliepimų duomenų šaltiniai: atsiliepimai apie Amazon parduotus mobilius telefonus [76], IMDB filmų įvertinimo duomenys [77] ir apsilankymo prie Eifelio bokšto Paryžiuje atsiliepimai, parašyti TripAdvisor svetainėje [78].

Ne visi duomenų rinkiniai buvo subalansuoti, taikomi metodai gauti subalansuotas klases: klasė, kuri turi daugiau stebinių, sumažinama iki mažesniosios klasės dydžio. Pradiniai duomenų rinkinių dydžiai ir po subalansavimo gautas duomenų pasiskirstymas klasėse pristatomas 4 lentelėje. Taip pat skiriasi vertinimo skalė: TripAdvisor ir Amazon vartotojai galėjo savo patirtį įvertinti nuo 1 iki 5. Šiais atvejais daroma prielaida, kad vartotojų, kurie įvertino savo patirtį 1-2 žvaigždutėmis sentimentas yra neigimas, 3 – neutralus, 4-5 – teigiamas. Kadangi IMDB galimos tik dvi klasės, nuspręsta visus duomenis klasifikuoti į dvi klases.

4 lentelė. Duomenų rinkiniai

Duomenys	**Skalė**	**Pradinio rinkinio dydis**	**Rinkinio dydis po subalansavimo ir tuščių reikšmių pašalinimo**	**Įrašų skaičius teigiamo sentimento klasėje**	**Įrašų skaičius neigiamo sentimento klasėje**
TripAdvisor	1-5	53077	27108	13554	13554
IMDB	0/1	50000	50000	25000	25000
Amazon	1-5	413840	194182	97104	97078

Kiekvienas duomenų rinkinys buvo nagrinėjamas atskirai: pirmiausiai iš visų atsiliepimų buvo padarytos leksemos, visos raidės paverstos mažosiomis, pašalinti skyrybos ženklai. Tuomet sudarytas žodynas – žodžiams priskirti unikalūs indeksai. Iš žodyno pašalinami žodžiai, kurie pasitaiko mažiau nei 3 dokumentuose. Tuomet sudaromas žodžių krepšelis – leksemoms priskiriami dažniai dokumentuose ir galiausiai apskaičiuojami TF-IDF indeksai.

5 lentelėje matoma, kad pradinis žodynas didžiausias IMDB duomenų aibėje, nors tai tik antras pagal dydį duomenų rinkinys. Iš to galima darysi prielaidą, kad filmus IMDB svetainėje

komentuojantys žmonės linkę rašyti ilgesnius atsiliepimus, naudoja įvairesnius žodžius, įdeda daugiau pastangų apibūdinti savo patirtį stebint mėgstamą arba labai nepatikusį filmą. Šiai prielaidai neprieštarauja paskaičiuota pradinio žodyno dalis, likusi po filtravimo – santykinė IMDB dalis procentais pati mažiausia, vadinasi, šios svetainės lankytojai naudoja įvairesnių ir ne taip dažnai naudojamų žodzių. didžiausia dalis pradinio žodyno liko TripAdvisor duomenyse, vadinasi šios svetainės lankytojai apie aplankytą Eifelio bokštą išsireiškia gana panašiais, dažnai vartojamais žodžiais.

5 lentelė. Žodyno sudarymas

Duomenys	**Pradinio žodyno dydis**	**Žodyno dydis atfiltravus retai pasitaikančius žodžius**	**Likusi pradinio žodyno dalis po filtravimo**
TripAdvisor	13601	8442	62.07%
IMDB	103136	48145	46.68%
Amazon	42158	22393	53.12%

Atsiliepimų transformavimui į vektorių buvo palyginti 6 metodai:

1. Doc2Vec_dbow _d128 – PV-DBOW metodas derinant žodžių ir dokumentų vektorius;
2. Doc2Vec_dbow_w _d128 – PV-DBOW metodas naudojant tik dokumentų vektorius;
3. Doc2Vec_dm_m_d128 – pastraipų vektorius – paskirstytos atminties modelis;
4. LSI – latentinis semantinis indeksavimas (TF-IDF vektoriaus transformacija)
5. RP – atsitiktinių projekcijų metodas (TF-IDF vektoriaus transformacija);
6. FastText – nustatytas 2-3 raidžių dydžio fiksuotas konteksto dydis.

Visiems vektorizavimo metodams buvo naudojamas fiksuotas vektoriaus ilgis, lygus 128 dimensijų. Korektiškam detekcijos gerumo įvertinimui buvo naudojama 10 dalių kryžminė patikra (angl. *cross-validation*).

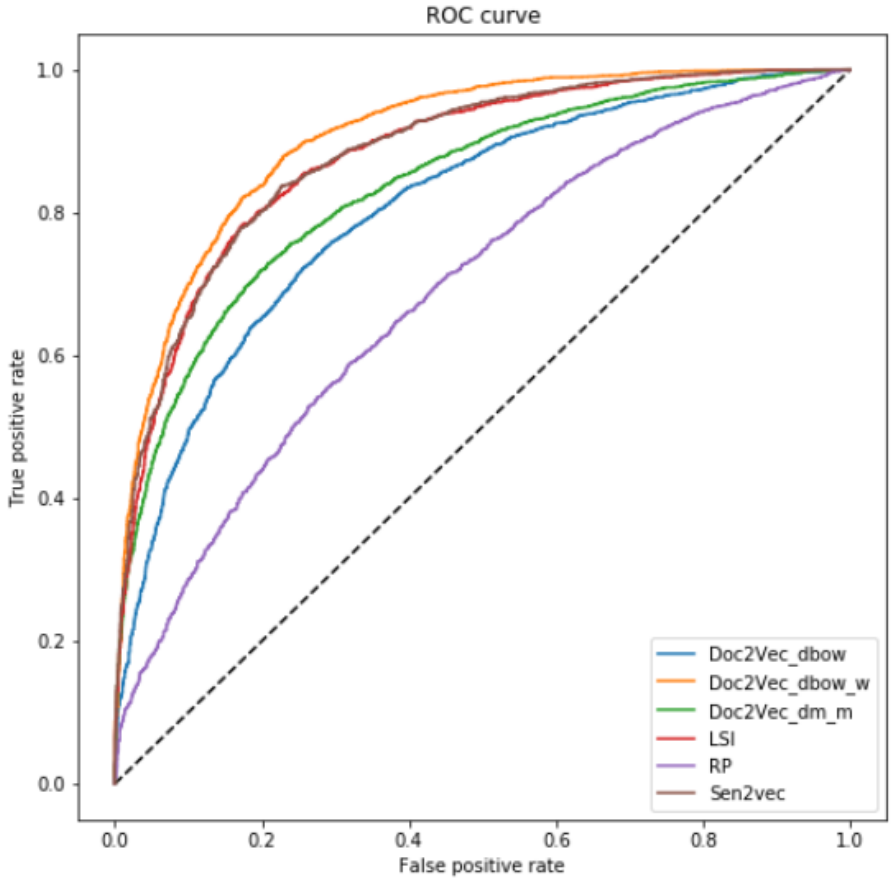
3.2.Sentimento detekcija su TripAdvisor duomenimis

Pritaikius 6 skirtingus vektorizavimo metodus TripAdvisor duomenimis ir su sudarytais vektorių rinkiniais išbandžius 3 skirtingus klasifikavimo algoritmus gauti rezultatai apibendrinami 6 lentelėje. Pagal AUC įvertį geriausiai pasirodė atsitiktinių projekcijų metodo ir atsitiktinio miško klasifikatoriaus kombinacija – pasiektas AUC = 0.98 rezultatas. Geriausiai logistinei regresijai ir neuroniniam tinklui tiko Doc2Vec_dbow_w _d128 metodu sukurti vektoriai, pasiektas rezultatas – 0.90. Įvertinus AUC įverčio vidurkį geriausią rezultatą iš vektorizavimo metodų, nepriklausomai nuo naudojamo detektoriaus, davė Doc2Vec_dbow_w _d128, LSI ir Sent2Vec. Geriausią vidutinį rezultatą tarp lyginamų metodų turi atsitiktinis miškas.

6 lentelė. Modelių AUC įverčių palyginimas su TripAdvisor duomenimis

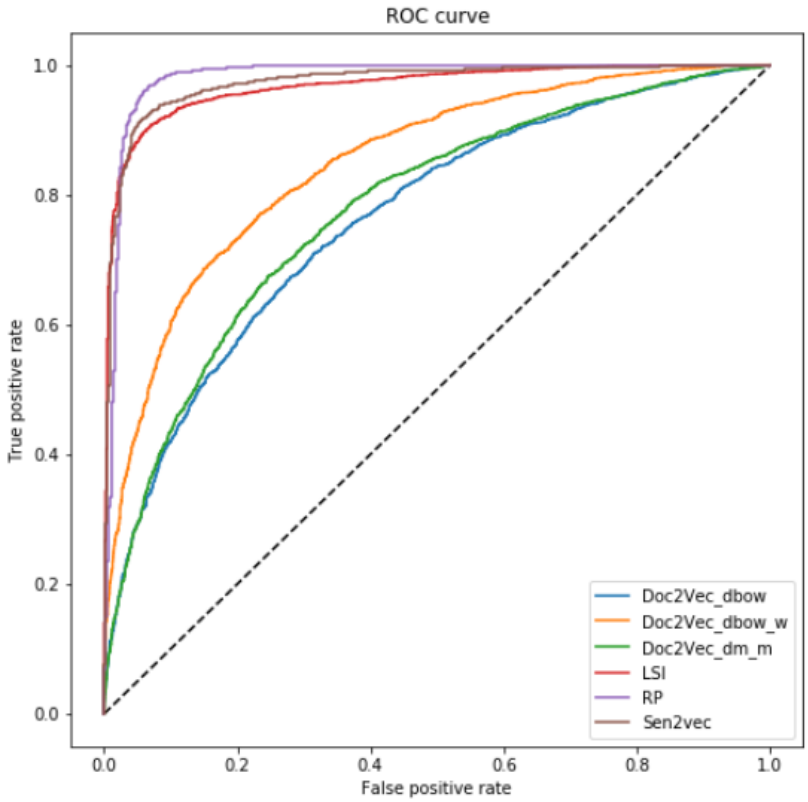
	Doc2Vec_dbow	Doc2Vec_dbow_w	Doc2Vec_dm_m	LSI	RP	Sent2Vec	Vidurkis
LR	0.80	**0.90**	0.84	0.88	0.69	0.88	0.83
RF	0.76	0.85	0.78	0.96	**0.98**	0.97	**0.88**
NN	0.80	**0.91**	0.83	0.88	0.70	0.88	0.83
Vidurkis	0.79	**0.89**	0.82	**0.91**	0.79	**0.91**	

Logistinės regresijos metodo ir visų vektorių sudarymui naudotų metodų rezultatams nubraižytos ROC kreivės pateikiamos 9 paveiksle. Pastebima, kad toliausiai nuo įstrižaines yra nutolusi Doc2Vec_dbow kreivė, vadinasi, šis modelis geriausiai klasifikuoja.



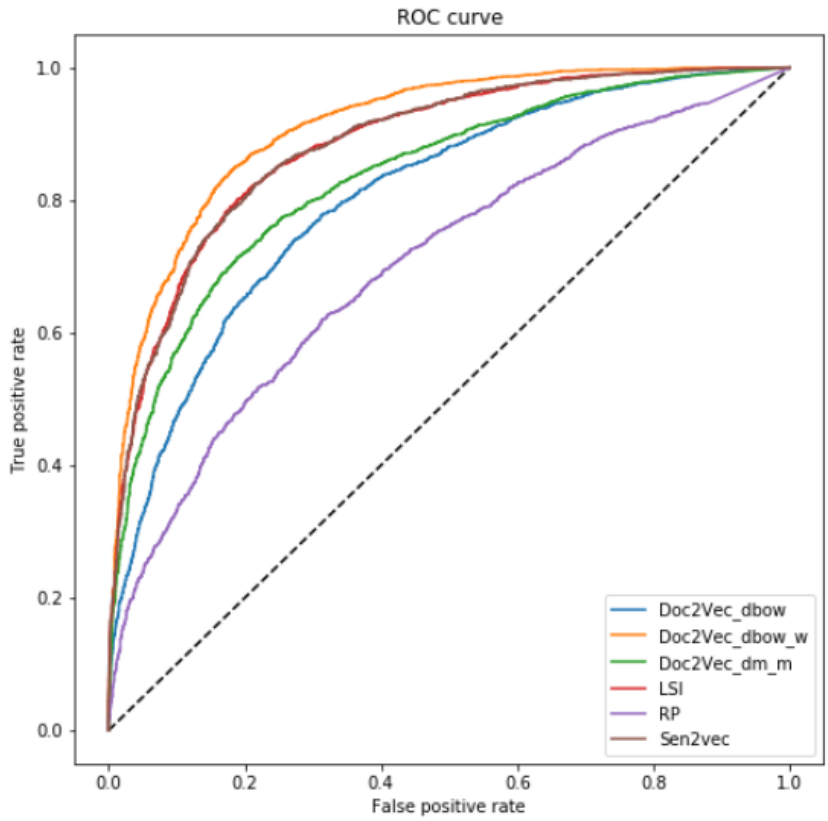
9 pav. Logistinės regresijos ROC kreivės su TripAdvisor duomenimis

Atsitiktinio miško ir visų vektorių sudarymui naudotų metodų rezultatams nubraižytos ROC kreivės pateikiamos 10 paveiksle.



10 pav. Atsitiktinio miško ROC kreivės su TripAdvisor duomenimis

Neuroninio tinklo ir visų vektorių sudarymui naudotų metodų rezultatams nubraižytos ROC kreivės pateikiamos 11 paveiksle.



11 pav. Neuroninio tinklo ROC kreivės su TripAdvisor duomenimis

Pateikiama geriausio modelio – atsitiktinio miško ir atsitiktinių projekcijų vektorizavimo kombinacijos sumaišymo matrica. Class 1 – teigiamas sentimentas, Class 2 – neigiamas sentimentas. Pastebima, kad klasės beveik simetriškos, todėl bendrasis tikslumas (91,66%) gali būti naudojamas modelio kokybei įvertinti. Taip pat, optimali atskyrimo tarp klasių riba lygi 0.4294729. Duomenų atskyrimo tarp klasių grafikai pateikiami 1 priede. Pagal tikslumo matą tiksliau prognozuojama teigiama klasė, tačiau iš teigiamų klasės teisingai atspėta tik 87,93%, todėl šiam modeliui būdinga antro tipo klaida.

		Truth data			
		Class 1	Class 2	Classification overall	Producer Accuracy (Precision)
Classifier results	Class 1	2572	99	2671	96.294%
	Class 2	353	2398	2751	87.168%
	Truth overall	2925	2497	5422	
	User Accuracy (Recall)	87.932%	96.035%		
Overall accuracy (OA):		91.664%			
Kappa[1]:		0.833			

12 pav. Geriausio modelio sumaišymo matrica su TripAdvisor duomenimis

Apibendrinus, nors tiksliausias modelis naudojo RP vektorizavimo metodą ir RF klasifikatorių – AUC įvertis lygus 98%, tačiau palyginus metodus pagal AUC įverčio vidurkį paaiškėjo, kad geriausiai vektorizuoja Doc2Vec_dbow_w _d128, LSI ir Sent2Vec.

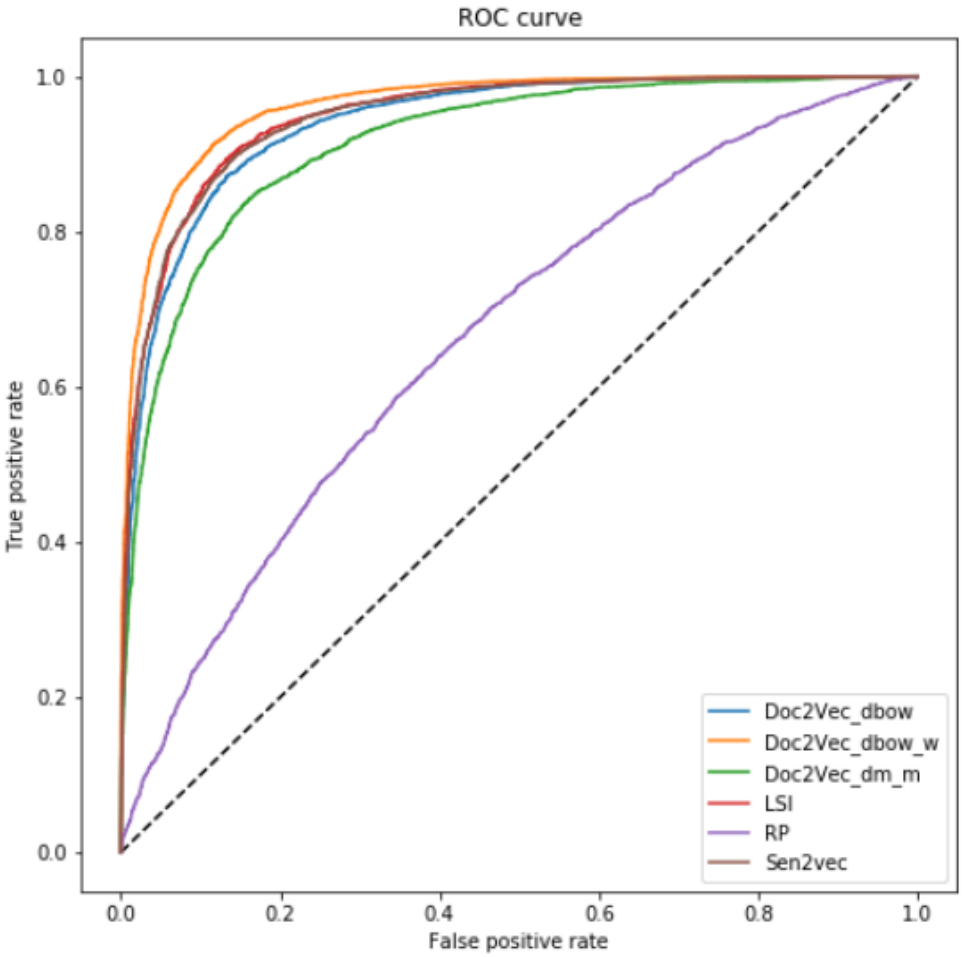
3.3.Sentimento detekcija su IMDB duomenimis

Pritaikius 6 skirtingus vektorizavimo metodus IMDB duomenimis ir su sudarytais vektorių rinkiniais išbandžius 3 skirtingus klasifikavimo algoritmus gauti rezultatai apibendrinami 7 lentelėje. Pagal AUC įvertį geriausiai pasirodė atsitiktinių Doc2Vec_dbow_w _d128 ir logistinės regresijos ir neuroninio tinklo klasifikatoriaus kombinacijos – pasiektas 0.96 rezultatas. Geriausiai atsitiktinių medžių klasifikatoriui tiko LSI metodu sukurti vektoriai, o pasiektas rezultatas labai panašus į geriausią LR ir NN rezultatą. Įvertinus AUC įverčio vidurkį geriausią rezultatą iš vektorizavimo metodų parodė Doc2Vec_dbow_w _d128, LSI ir Sent2Vec. Geriausią rezultatą iš klasifikavimo metodų parodė logistinė regresija.

7 lentelė. Modelių AUC įverčių palyginimas su IMDB duomenimis

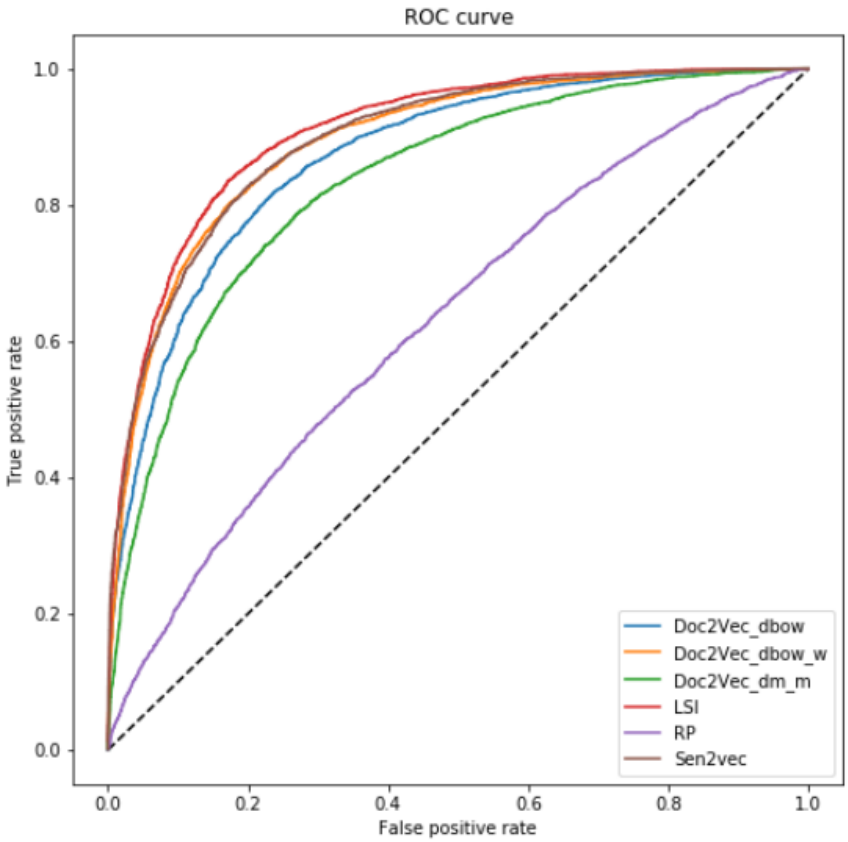
	Doc2Vec_dbow	Doc2Vec_dbow_w	Doc2Vec_dm_m	LSI	RP	Sent2Vec	Vidurkis
LR	0.94	**0.96**	0.92	0.95	0.67	0.95	**0.90**
RF	0.87	0.89	0.83	**0.95**	0.63	0.90	0.85
NN	0.94	**0.95**	0.82	0.91	0.66	**0.95**	0.87
Vidurkis	0.92	**0.94**	0.86	**0.94**	0.65	**0.93**	

Logistinės regresijos metodo ir visų vektorių sudarymui naudotų metodų rezultatams nubraižytos ROC kreivės pateikiamos 13 paveiksle.



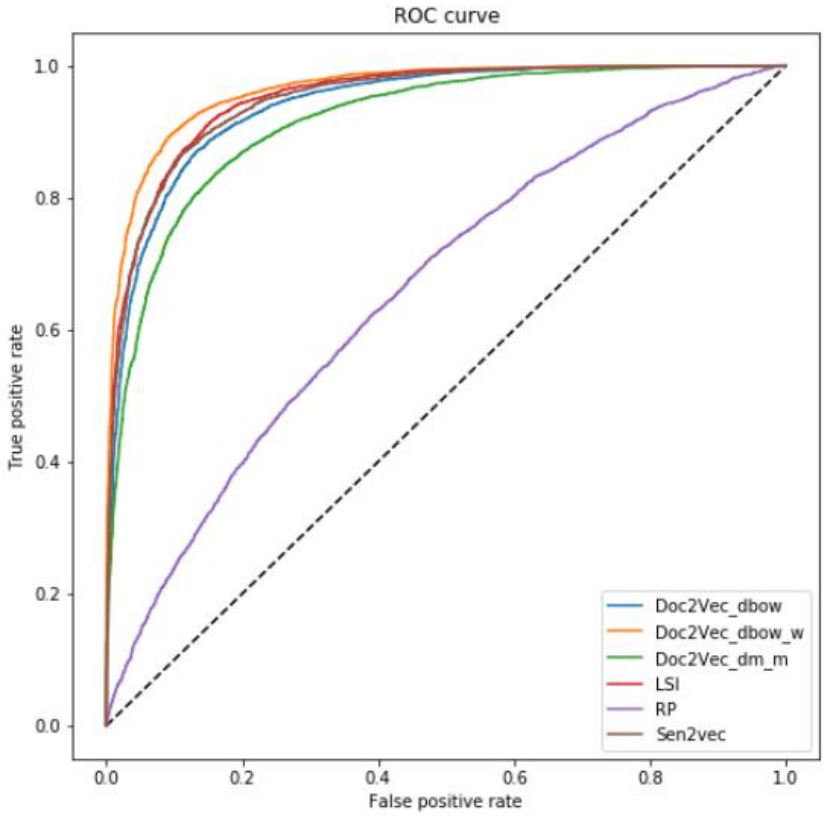
13 pav. Logistinės regresijos ROC kreivės su IMDB duomenimis

Atsitiktinio miško ir visų vektorių sudarymui naudotų metodų rezultatams nubraižytos ROC kreivės pateikiamos 14 paveiksle.



14 pav. Atsitiktinio miško ROC kreivės su IMDB duomenimis

Neuroninio tinklo ir visų vektorių sudarymui naudotų metodų rezultatams nubraižytos ROC kreivės pateikiamos 15 paveiksle.



15 pav. Neuroninio tinklo ROC kreivės su IMDB duomenimis

Pateikiama geriausio modelio – atsitiktinio miško ir LSI kombinacijos sumaišymo matrica. Class 1 – teigiamas sentimentas, Class 2 – neigiamas sentimentas. Pastebima, kad klasės beveik simetriškos, todėl bendrasis tikslumas (83%) gali būti naudojamas modelio kokybei įvertinti. Taip pat, optimali atskyrimo tarp klasių riba lygi 0.5154003. Duomenų atskyrimo tarp klasių grafikai pateikiami 1 priede. Pagal tikslumo matą tiksliau prognozuojama neigiama klasė, tačiau iš teigiamų klasės teisingai atspėta tik 81,77%, todėl šiam modeliui būdinga pirmo tipo klaida.

		Truth data			
		Class 1	Class 2	Classification overall	Producer Accuracy (Precision)
Classifier results	Class 1	4124	931	5055	81.583%
	Class 2	769	4176	4945	84.449%
	Truth overall	4893	5107	10000	
	User Accuracy (Recall)	84.284%	81.77%		

Overall accuracy (OA):	83%
Kappa[1]:	0.66

Apibendrinus, nors tiksliausi modeliai naudojo Doc2Vec_dbow_w_d128 vektorizavimo metodą ir LR, NN klasifikatorius – AUC įvertis lygus 96%, tačiau palyginus metodus pagal AUC įverčio vidurkį paaiškėjo, kad geriausiai vektorizuoja Doc2Vec_dbow_w_d128, LSI metodas.

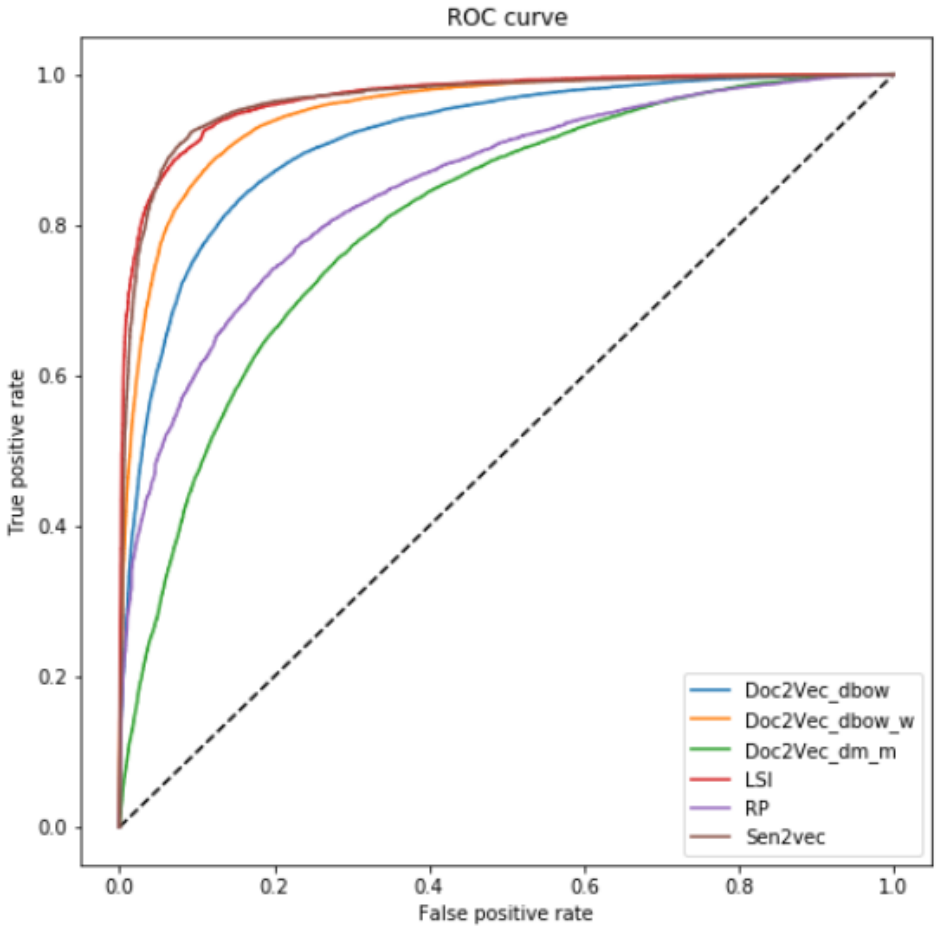
3.4.Sentimento detekcija su Amazon duomenimis

Pritaikius 6 skirtingus vektorizavimo metodus Amazon duomenimis ir su sudarytais vektorių rinkiniais išbandžius 3 skirtingus klasifikavimo algoritmus gauti rezultatai apibendrinami 8 lentelėje. Pagal AUC įvertį geriausiai pasirodė Sent2Vec ir atsitiktinių medžių klasifikatoriaus kombinacija – pasiektas 0.99 rezultatas. Geriausiai logistinei regresijai ir neuroniniam tinklui taip pat tiko Sent2Vec metodu sukurti vektoriai, tačiau pasiektas rezultatas –0.97 atsilieka nuo RF. Įvertinus AUC įverčio vidurkį geriausią rezultatą iš vektorizavimo metodų parodė Sent2Vec – 0.98. Geriausią rezultatą iš klasifikavimo metodų parodė neuroninis tinklas – 0.93.

8 lentelė. Modelių AUC įverčių palyginimas su Amazon duomenimis

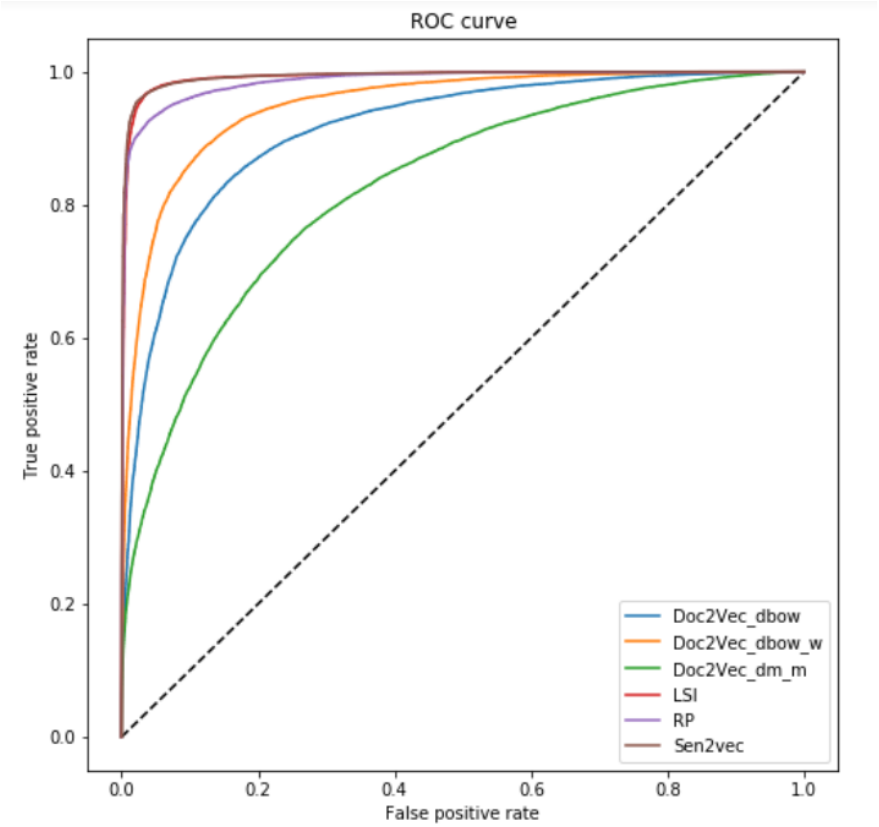
	Doc2Vec_dbow	Doc2Vec_dbow_w	Doc2Vec_dm_m	LSI	RP	Sent2Vec	Vidurkis
LR	0.91	0.95	0.81	0.91	0.85	**0.97**	0.90
RF	0.89	0.94	0.83	0.99	0.94	**0.99**	0.91
NN	0.93	0.96	0.86	0.97	0.87	**0.97**	**0.93**
Vidurkis	0.91	**0.95**	0.83	**0.96**	0.89	**0.98**	

Logistinės regresijos metodo ir visų vektorių sudarymui naudotų metodų rezultatams nubraižytos ROC kreivės pateikiamos 16 paveiksle.



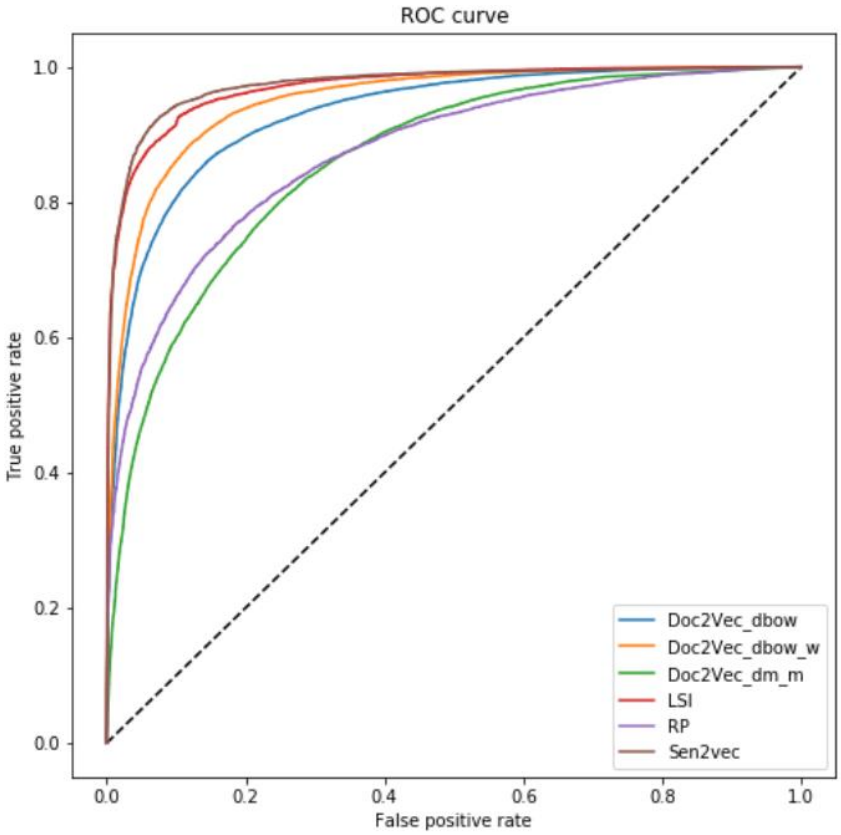
16 pav. Logistinės regresijos ROC kreivės su IMDB duomenimis

Atsitiktinio miško ir visų vektorių sudarymui naudotų metodų rezultatams nubraižytos ROC kreivės pateikiamos 17 paveiksle.



17 pav. Atsitiktinio miško ROC kreivės su IMDB duomenimis

Neuroninio tinklo ir visų vektorių sudarymui naudotų metodų rezultatams nubraižytos ROC kreivės pateikiamos 18 paveiksle.



18 pav. Neuroninio tinklo ROC kreivės su IMDB duomenimis

Pateikiama geriausio modelio – atsitiktinio miško ir Sen2Vec vektorizavimo kombinacijos sumaišymo matrica. Class 1 – teigiamas sentimentas, Class 2 – neigiamas sentimentas. Pastebima, kad klasės beveik simetriškos, todėl bendrasis tikslumas (96,56%) gali būti naudojamas modelio kokybei įvertinti. Taip pat, optimali atskyrimo tarp klasių riba lygi 0.6602915. Duomenų atskyrimo tarp klasių grafikai pateikiami 1 priede. Pagal tikslumo matą tiksliau prognozuojama teigiama klasė.

		Truth data			
		Class 1	Class 2	Classification overall	Producer Accuracy (Precision)
Classifier results	Class 1	18782	525	19307	97.281%
	Class 2	811	18707	19518	95.845%
	Truth overall	19593	19232	38825	
	User Accuracy (Recall)	95.861%	97.27%		
Overall accuracy (OA):		96.559%			
Kappa[1]:		0.931			

19 pav. RF + Sen2Vec ROC kreivės su IMDB duomenimis

Apibendrinus, Amazon buvo vienintelis duomenų rinkinys, su kuriuo Sent2Vec vektorizavimo metodas ir LR, RF ir NN klasifikatoriai turėjo aukščiausią AUC įvertį, atitinkamai lygų 97%, 99%, 97%. Taigi palyginus metodus pagal AUC įverčio vidurkį paaiškėjo, kad geriausiai vektorizuoja Sent2Vec, o LSI ir Doc2Vec_dbow_w rezultatas labai panašus. Kadangi visiems duomenų rinkiniams geriausią vidutinį rezultatą tarp lyginamų metodų turi atsitiktinis miškas, toliau tyrime iš mašininio mokymo metodų buvo nagrinėjamos tik atsitiktinio miško kombinacijos su vektorizavimo metodais.

3.5.Sentimento detekcija žodynu grįstais metodais

Siekiant palyginti mašininio mokymosi algoritmus su grįstais žodynu, buvo atlikta sentimentų detekcija su tais pačiais duomenų rinkiniais naudojant programinės įrangos R paketą SemanticAnalysis ir jame pateikiamus žodynus:

- GI = Harvard-IV žodynas
- HE = Henry's finansų srities žodynas (Journal of Business Communication)
- LM = Loughran-McDonald finansų srities žodynas
- QDAP = sentimentų žodynas

Gauti rezultatai pateikiami naudojant AUC metriką, ROC kreives ir koreliacijos matricas.

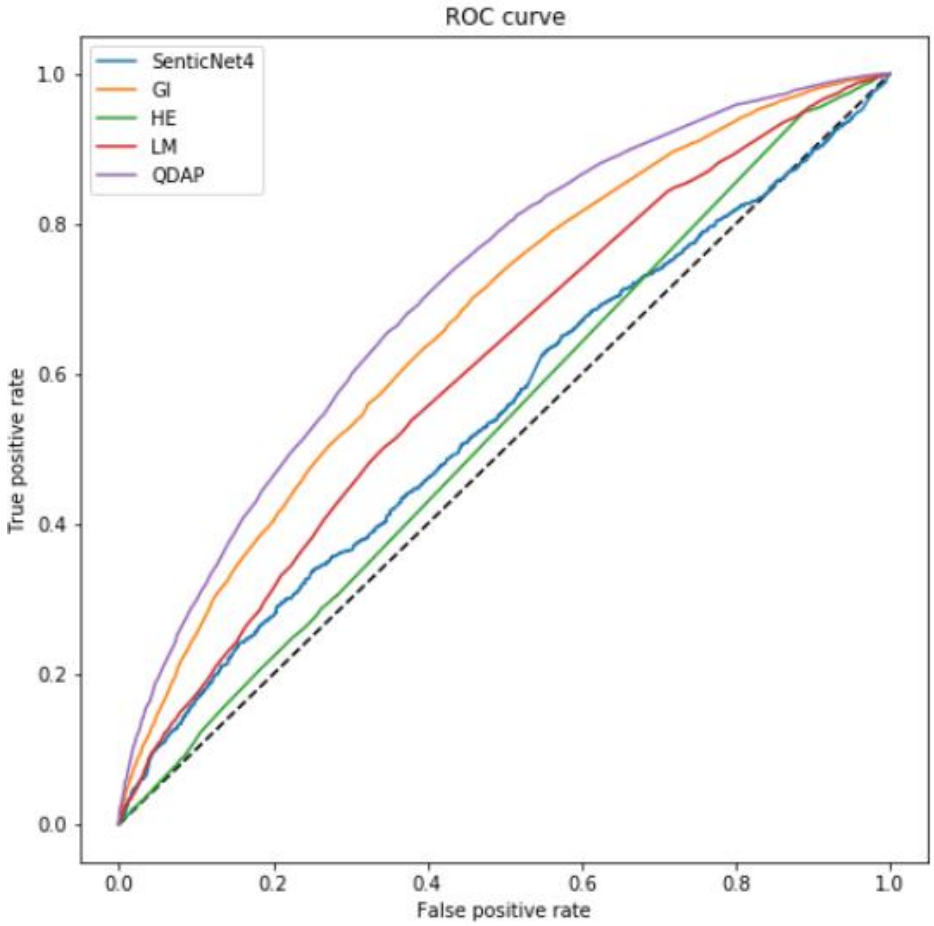
Naudojant TripAdvisor duomenų rinkinį didžiausias tikslumas pasiekiamas su QDAP žodynu – 0.7128. Tai patvirtina ir ROC kreivių palyginimas.

9 lentelė. Žodynu grįstų modelių AUC įverčių palyginimas su TripAdvisor duomenimis

10 lentelė. TripAdvisor AUC metrika

TripAdvisor	**AUC**
SentimentGI	0.6718
SentimentHE	0.5327
SentimentLM	0.6087
SentimentQDAP	0.7128
SenticNet4	0.5468

.



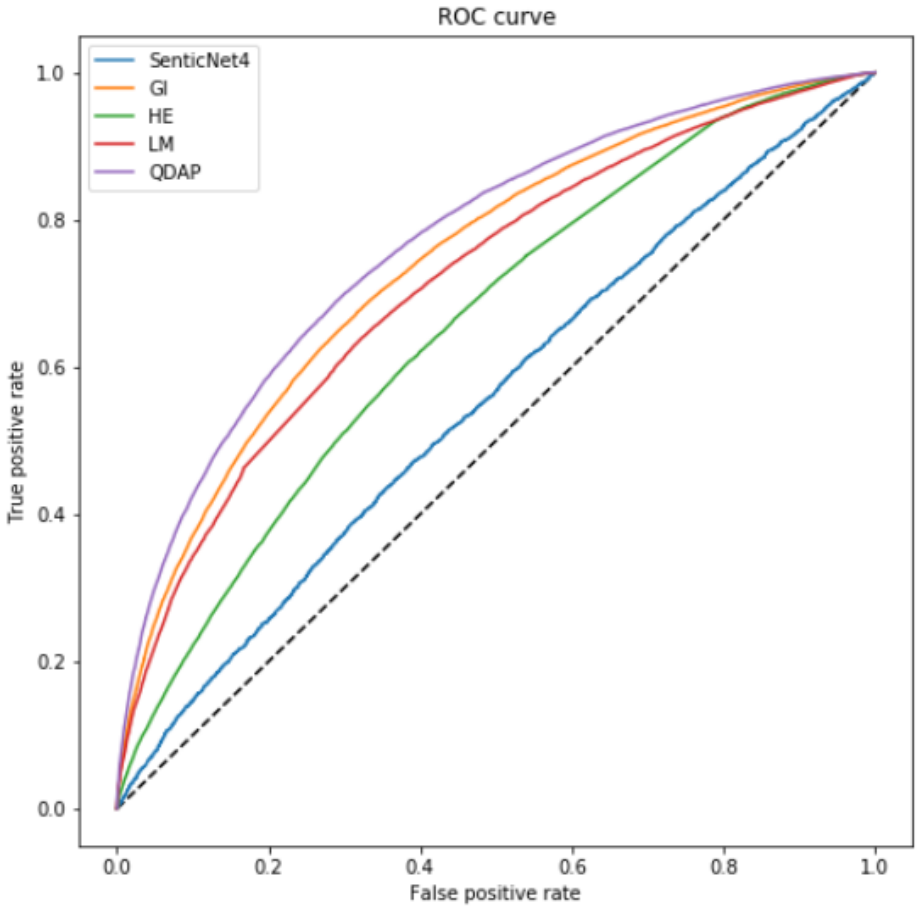
20 pav. Žodynu grįstų metodų ROC kreivės Su TripAdvisor duomenimis

Naudojant IMDB duomenų rinkinį didžiausias tikslumas taip pat pasiekiamas su QDAP žodynu – 0.7691. Tai patvirtina ir ROC kreivių palyginimas.

11 lentelė. IMDB AUC metrika

IMDB	**AUC**
SentimentGI	0.7422
SentimentHE	0.6554
SentimentLM	0.7149
SentimentQDAP	0.7691

SenticNet4 0. 5519

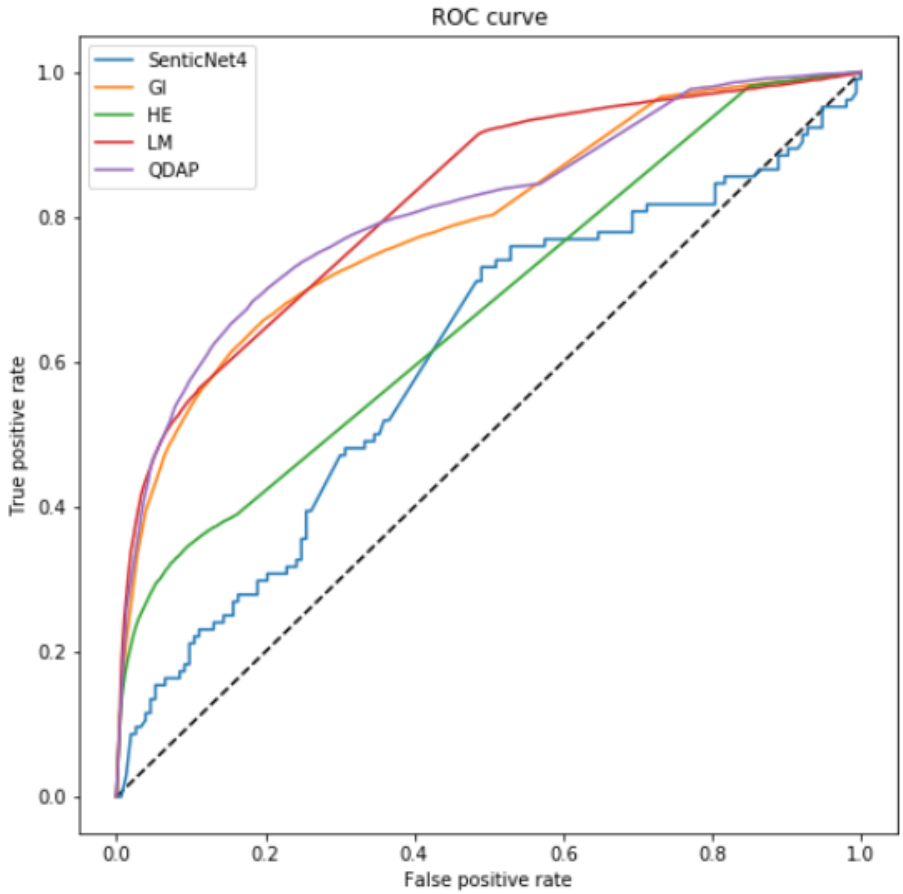


21 pav. Žodynu grįstų metodų ROC kreivės Su IMDB duomenimis

Naudojant Amazon duomenų rinkinį didžiausias tikslumas pasiekiamas su LM žodynu – 0.8195, tačiau tai labai artimas rezultatas QDAP metodui. Tai patvirtina ir ROC kreivių palyginimas.

12 lentelė. Amazoon AUC metrika

Amazon	**AUC**
SentimentGI	0.7904
SentimentHE	0.6694
SentimentLM	0.8195
SentimentQDAP	0.8094
SenticNet4	0.6049



22 pav. Žodynu grįstų metodų ROC kreivės Su Amazon duomenimis

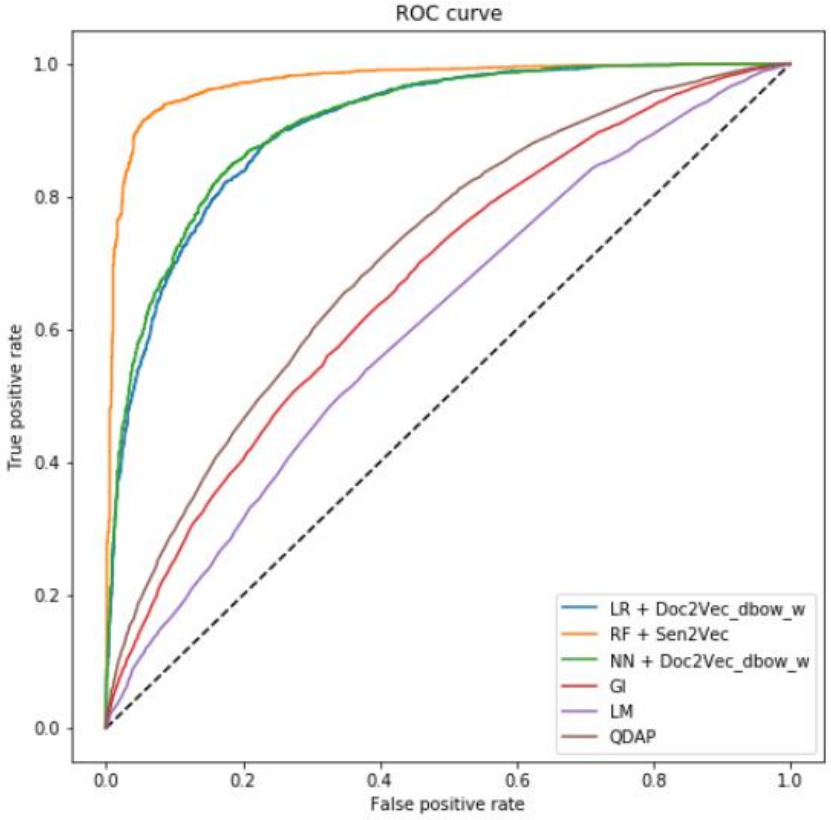
Apibendrinant, naudojant žodynu grįstus metodus geriausias rezultatas buvo pasiektas su Amazon duomenimis ir LM metodu, tačiau atsižvelgiant į metodų veikimą su visais duomenų rinkiniais, QDAP metodas atrodo stabiliausias ir rodantis geriausius rezultatus.

3.6.Hibridinio modelio kūrimas

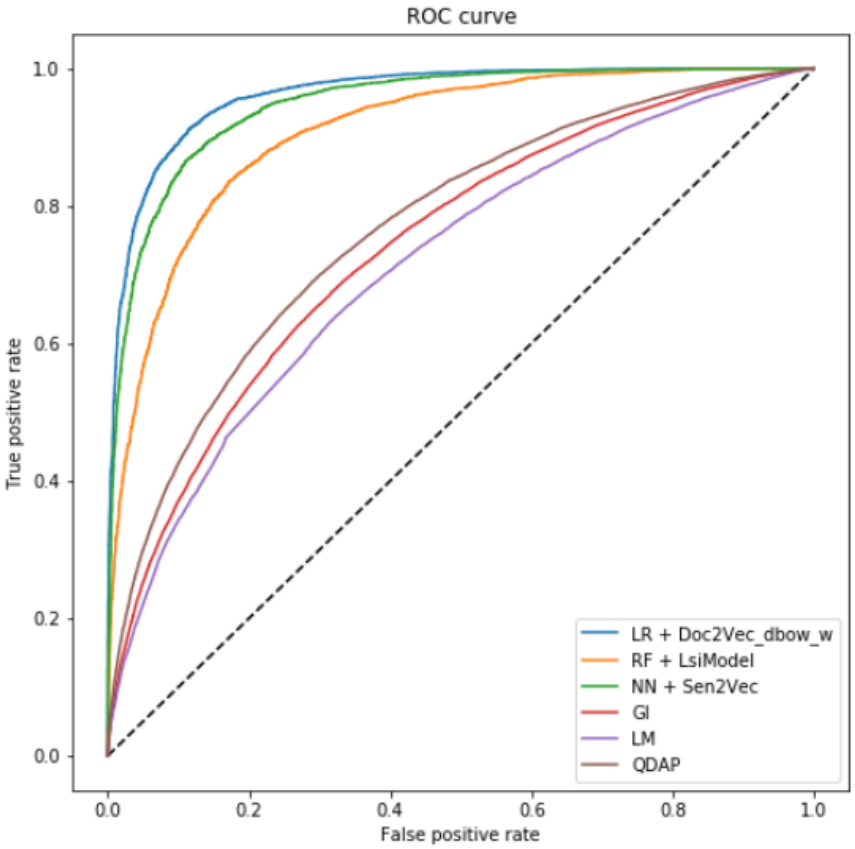
Hibridinis modelis – kelių modelių apjungimas vykdomas siekiant sukurti tikslesnį klasifikavimo modelį nei taikant tuos modelius atskirai. Šiame tyrime buvo pasirinkta sujungti kiekvienam duomenų rinkiniui po 3 geriausius modelius iš sudarytų mašininiu mokymu ir žodynu grįstų metodų.

Žemiau pateikiami atrinktų geriausių modelių palyginimo ROC kreivių grafikai. Pastebėta, kad mašininio mokymo metodai su visais duomenų rinkiniais prognozuoja tiksliau metodai, grįsti žodynu. Su TripAdvisor duomenimis iš pasirinktų modelių tiksliausiai sentimentus klasifikavo atsitiktinio miško algoritmas kartu su Doc2Vec_dbow_w metodu (žr. 16 pav.). Nors pradinėje analizėje geriausiai pasirodė atsitiktiniai miškai su atsitiktinių projekcijų metodu RP, tačiau RP pasirodė prastai su visais kitais metodais, tad šįkart jis išrinktas tarp geriausių nebuvo. Su IMDB duomenimis iš pasirinktų modelių tiksliausiai sentimentus klasifikavo logistinė regresija kartu su Doc2Vec_dbow_w ir neuroniniai tinklai kartu su LSI metodu (žr. 17 pav.). Su Amazon duomenimis iš pasirinktų modelių

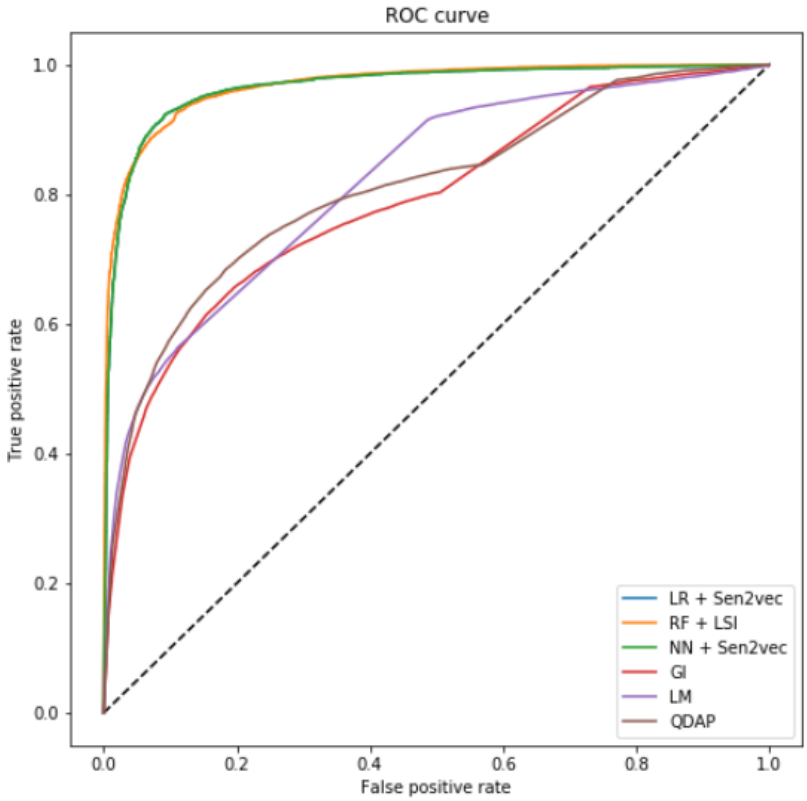
tiksliausiai sentimentus klasifikavo atsitiktinio miško algoritmas kartu su Sen2Vec metodu (žr. 23 pav.).



23 pav. Geriausi TripAdvisor klasifikavimo metodai



24 pav. Geriausi IMDB klasifikavimo metodai



25 pav. Geriausi Amazon klasifikavimo metodai

Atrinkus aukščiau aptartis geriausius modelius gaunamas iš 6 modelių sudarytas hibridinis modelis. Panaudojamas apmokymas ant visų duomenų Visiems duomenų rinkiniams sukurti modeliai ir AUC metrika įvertintas jų tikslumas pateikti 10 lentelėje. Matoma, kad su sujungtais modeliais pasiektas geriausias AUC rezultatas visiems duomenų rinkiniams. Taip pat pateikiamos klasių atskyrimo reikšmės taške, kuriame jautrumas lygus specifiškumui.

13 lentelė. Hibridiniai modeliai

Duomenų rinkinys	**Sujungti modeliai**		**AUC**	**Ribinė reikšmė**
	Mašininiu mokymu grįsti	**Žodynu grįsti**		
TripAdvisor	LR + Doc2Vec_dbow_w RF + Sent2Vec NN + Doc2Vec_dbow_w	GI LM QDAP	0.99981	0.67543
IMDB	LR + Doc2Vec_dbow_w RF + LSI NN + Sent2Vec	GI LM QDAP	0.99979	0.71126
Amazon	LR + Sent2Vec RF + Sent2Vec NN + Sent2Vec	GI LM QDAP	0.99965	0.67029

Taigi sujungus geriausius modelius sentimentų klasifikavimas tampa itin tikslus. Visiems duomenų rinkiniams AUC metrikos reikšmė didelė – tai reškia, kad klasifikavimo kokybė pagerėjo ir prognozuojant sentimento poliariškumą pasitaiko labai mažai klaidų.

Išvados

1. Išanalizavus mokslinę literatūrą buvo išsiaiškinta, kad naudodama automatizuotą sentimentų analizę kompanija gali padidinti vartotojo pasitenkinimą ir pagerinti vartotojo patirtį taip įgaudama konkurencinį pranašumą. Sentimentų analizė taip pat įgalina sekti vartotojų sentimentą realiuoju laiku, tad kompanija gali imtis veiksmų vos tik pastebėjus reikšmingus pokyčius vartotojų atsiliepimuose.
2. Sentimentų analizėje naudojami dviejų grupių metodai – pagrįsti mašininiu mokymu arba žodynais. Kadangi mašininiu mokymu paremtiems metodams nereikia sudaryti sentimentų žodyno, taip pat išsprendžiama žodžių tvarkos, konteksto ir semantinės teksto prasmės problema, todėl šie metodai itin dažnai naudojamai, kai turimas didelis duomenų rinkinys. Šiame tyrime nuspręsta palyginti mašininiu mokymu ir žodynu grįstus metodus ir sujungti geriausius modelius.
3. Palyginus klasifikavimo metodus pastebėta, kad atsitiktiniai miškai parodė geriausius rezultatus vertinant pagal AUC įvertį ir ROC kreivę. Iš vektorių sudarymo metodų geriausius rezultatus parode Sen2Vec ir latentinio semantinio indeksavimo metodai.
4. Iš taikytų metodų atrinkti 3 geriausi mašininiu mokymu grįsti atsitiktinio miško ir skirtingos vektorizavimo metodų kombinacijos buvo sujungti su žodynais grįstais metodais ir gautas AUC rezultatas buvo geriausias iš visų atskirai taikytų tų pačių klasifikavimo metodų. Nors pradinių modelių tikslumas buvo aukštas, klasifikavimo kokybė sujungus kelis metodus padidėjo.

Literatūros sąrašas

[1] H. Akkineni, P. V. S. Lakshmi, and B. Vijay Babu, "Online Crowds Opinion-Mining it to Analyze Current Trend: A Review," *Int. J. Electr. Comput. Eng. J.*, vol. 5, no. 5, pp. 2088–8708, 2015.

[2] S. R. Das *et al.*, "Yahoo! for amazon: Sentiment extraction from small talk on the web," *8TH ASIA PACIFIC Financ. Assoc. Annu. Conf.*, 2001.

[3] T. Nasukawa and J. Yi, "Sentiment analysis: Capturing favorability using natural language processing," *Proc. Int. Conf. Knowl. capture - K-CAP '03*, no. January, p. 70, 2003.

[4] E. Cambria, D. Das, S. Bandyopadhyay, and A. (Editors) Feraco, *A Practical Guide to Sentiment Analysis (Socio-Affecting Computing 5)*. 2017.

[5] B. Liu, *Sentiment Analysis and Opinion Mining*, no. May. 2012.

[6] G. Emil and W. Salib, "Measuring the effect of Viral Negative Sentiment on Market Value: Case Study on United Airlines Crisis 2017," 2017.

[7] Z.-P. Fan, Y.-J. Che, and Z.-Y. Chen, "Product sales forecasting using online reviews and historical sales data: A method combining the Bass model and sentiment analysis," *J. Bus. Res.*, vol. 74, pp. 90–100, May 2017.

[8] S. W. K. Chan and M. W. C. Chong, "Sentiment analysis in financial texts," *Decis. Support Syst.*, vol. 94, pp. 53–64, Feb. 2017.

[9] M. Daniel, R. F. Neves, and N. Horta, "Company event popularity for financial markets using Twitter and sentiment analysis," *Expert Syst. Appl.*, vol. 71, pp. 111–124, Apr. 2017.

[10] K. Cortis *et al.*, "SemEval-2017 Task 5: Fine-Grained Sentiment Analysis on Financial Microblogs and News," pp. 519–535, 2017.

[11] G. Zhang, L. Xu, and Y. Xue, "Model and forecast stock market behavior integrating investor sentiment analysis and transaction data," *Cluster Comput.*, vol. 20, no. 1, pp. 789–803, Mar. 2017.

[12] J. Wu, "Review popularity and review helpfulness: A model for user review effectiveness," *Decis. Support Syst.*, vol. 97, pp. 92–103, 2017.

[13] S. Karimi and F. Wang, "Online review helpfulness: Impact of reviewer profile image," *Decis. Support Syst.*, vol. 96, pp. 39–48, 2017.

[14] M. Etter, E. Colleoni, L. Illia, K. Meggiorin, and A. D 'eugenio, "Measuring Organizational Legitimacy in Social Media: Assessing Citizens' Judgments With Sentiment Analysis," *Bus. Soc.*, vol. 57, no. 1, pp. 60–97, 2018.

[15] A. R. Alaei, S. Becken, and B. Stantic, "Sentiment Analysis in Tourism: Capitalizing on Big Data," *J. Travel Res.*, p. 4728751774775, 2017.

[16] Z. Xu, H. Zhang, V. Sugumaran, K.-K. Raymond Choo, L. Mei, and Y. Zhu, "Participatory

sensing-based semantic and spatial analysis of urban emergency events using mobile social media," 2011.

[17] K. Wegba, A. Lu, Y. Li, and W. Wang, "Interactive Movie Recommendation Through Latent Semantic Analysis and Storytelling," 2017.

[18] P. W. Farris, N. T. Bendle, P. E. Pfeifer, and D. J. Reibstein, "THE DEFINITIVE GUIDE TO MEASURING MARKETING PERFORMANCE."

[19] D. Kang and Y. Park, "Review-based measurement of customer satisfaction in mobile service: Sentiment analysis and VIKOR approach," *Expert Syst. Appl.*, vol. 41, no. 4, pp. 1041–1050, Mar. 2014.

[20] N. Hill and J. Brierley, *How to Measure Customer Satisfaction*. Routledge, 2017.

[21] R. Agnihotri, R. Dingus, M. Y. Hu, and M. T. Krush, "Social media: Influencing customer satisfaction in B2B sales," 2016.

[22] A. Sarkar Sengupta, M. S. Balaji, and B. C. Krishnan, "How customers cope with service failure? A study of brand reputation and customer satisfaction," *J. Bus. Res.*, vol. 68, no. 3, pp. 665–674, Mar. 2015.

[23] S. Chheda, E. Duncan, and S. Roggenhofer, "Putting customer experience at the heart of next-generation operating models," *Digit. McKinsey*, 2017.

[24] T. H. Engler, P. Winter, and M. Schulz, "Understanding online product ratings: A customer satisfaction model," *J. Retail. Consum. Serv.*, vol. 27, pp. 113–120, 2015.

[25] U. Ramanathan, N. Subramanian, and G. Parrott, "Role of social media in retail network operations and marketing to enhance customer satisfaction," *Int. J. Oper. Prod. Manag.*, vol. 37, no. 1, pp. 105–123, Jan. 2017.

[26] Q. Nguyen, T. Knox, and D. Prabhakar, "Understanding customer satisfaction in the UK quick service restaurant in-dustry: The influence of the tangible attributes of perceived service quality.," *Br. Food J.*, vol. 103, no. 1, pp. 36–45.

[27] C. Fornell, F. V. Morgeson, and G. T. M. Hult, "Stock Returns on Customer Satisfaction Do Beat the Market: Gauging the Effect of a Marketing Intangible," *J. Mark.*, vol. 80, no. 5, pp. 92–107, 2016.

[28] L. Abbott, *Quality and competition;an essay in economic theory*. New York :, 1955.

[29] W. Alderson, *Marketing behavior and executive action : a functionalist approach to marketing theory*. Martino Pub, 2009.

[30] B. J. Pine and J. H. Gilmore, *The Experience Economy: Work is Theater & Every Business a Stage*, vol. 76, no. 4. Harvard Business School Press, 1999.

[31] B. Schmitt, J. Joško Brakus, and L. Zarantonello, "From experiential psychology to consumer experience," *J. Consum. Psychol.*, vol. 25, no. 1, pp. 166–171, Jan. 2015.

[32] R. Bolton *et al.*, “Small Details that Make Big Differences: a radical approach to consumption experience as a firm’s differentiating strategy,” 2014.

[33] C. Meyer and A. Schwager, “Understanding Customer Experience,” *Harv. Bus. Rev.*, vol. 85, no. 2, pp. 117–26, 2007.

[34] S. Hedegaard and J. G. Simonsen, “Extracting usability and user experience information from online user reviews,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI ’13*, 2013, p. 2089.

[35] A.-M. Kranzbühler, M. H. P. Kleijnen, R. E. Morgan, and M. Teerling, “The Multilevel Nature of Customer Experience Research: An Integrative Review and Research Agenda,” *Int. J. Manag. Rev.*, vol. 20, no. 2, pp. 433–456, Apr. 2018.

[36] Mccoll-Kennedy, “Fresh perspectives on customer experience,” *J. Serv. Mark.*, vol. 29, pp. 6–7, 2015.

[37] J. R. Mccoll-Kennedy and L. Cheung, “CO-CREATING SERVICE EXPERIENCE PRACTICES,” *J. Serv. Manag.*, 2015.

[38] K. Bauman, B. Liu, and A. Tuzhilin, “Recommending Items with Conditions Enhancing User Experiences Based on Sentiment Analysis of Reviews,” *CBRecSys@ RecSys*, 2016.

[39] P. Foroudi, S. Gupta, U. Sivarajah, and A. Broderick, “Investigating the effects of smart technology on customer dynamics and customer experience,” *Comput. Human Behav.*, vol. 80, pp. 271–282, Mar. 2018.

[40] P. C. Verhoef, K. N. Lemon, A. Parasuraman, A. Roggeveen, M. Tsiros, and L. A. Schlesinger, “Customer Experience Creation: Determinants, Dynamics and Management Strategies,” *J. Retail.*, vol. 85, no. 1, pp. 31–41, 2009.

[41] S.-H. Chang, W.-H. Chih, D.-K. Liou, and Y.-T. Yang, “The mediation of cognitive attitude for online shopping,” *Inf. Technol. People*, vol. 29, no. 3, pp. 618–646, Aug. 2016.

[42] F. H. Khan, U. Qamar, and S. Bashir, “eSAP: A decision support framework for enhanced sentiment analysis and polarity classification,” *Inf. Sci. (Ny).*, vol. 367–368, pp. 862–873, Nov. 2016.

[43] A. Collomb, C. Costea, D. Joyeux, O. Hasan, and L. Brunie, “A Study and Comparison of Sentiment Analysis Methods for Reputation Evaluation.”

[44] S. Sun, C. Luo, and J. Chen, “A review of natural language processing techniques for opinion mining systems,” *Inf. Fusion*, vol. 36, pp. 10–25, 2017.

[45] F. N. Ribeiro, M. Araújo, P. Gonçalves, M. André Gonçalves, and F. Benevenuto, “SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods,” *EPJ Data Sci.*, vol. 5, no. 1, p. 23, Dec. 2016.

[46] E. Cambria, S. Poria, D. Hazarika, and K. Kwok, “SenticNet 5: Discovering Conceptual

Primitives for Sentiment Analysis by Means of Context Embeddings," in *AAAI*, 2018.

[47] G. Markopoulos, G. Mikros, A. Iliadi, and M. Liontos, "Sentiment Analysis of Hotel Reviews in Greek: A Comparison of Unigram Features," *Cult. Tour. a Digit. Era*, 2015.

[48] L. Zheng, H. Wang, and S. Gao, "Sentimental feature selection for sentiment analysis of Chinese online reviews," *Int. J. Mach. Learn. Cybern.*, vol. 9, no. 1, pp. 75–84, Jan. 2018.

[49] C. Pujari, Aiswarya, and N. P. Shetty, "Comparison of Classification Techniques for Feature Oriented Sentiment Analysis of Product Review Data," Springer, Singapore, 2018, pp. 149–158.

[50] Z. Xiang, Z. Schwartz, J. H. Gerdes, and M. Uysal, "What can big data and text analytics tell us about hotel guest experience and satisfaction?," *Int. J. Hosp. Manag.*, vol. 44, pp. 120–130, Jan. 2015.

[51] M. Rossetti, F. Stella, and M. Zanker, "Analyzing user reviews in tourism with topic models," *Inf. Technol. Tour.*, vol. 16, no. 1, pp. 5–21, Mar. 2016.

[52] C. Lin and Y. He, "Joint sentiment/topic model for sentiment analysis," in *Proceeding of the 18th ACM conference on Information and knowledge management - CIKM '09*, 2009, p. 375.

[53] M. Giatsoglou, M. G. Vozalis, K. Diamantaras, A. Vakali, G. Sarigiannidis, and K. C. Chatzisavvas, "Sentiment analysis leveraging emotions and word embeddings," *Expert Syst. Appl.*, vol. 69, pp. 214–224, Mar. 2017.

[54] H. Saif, Y. He, M. Fernandez, and H. Alani, "Contextual semantics for sentiment analysis of Twitter," *Inf. Process. Manag.*, vol. 52, no. 1, pp. 5–19, Jan. 2016.

[55] D. Chatzakou, V. Koutsonikola, A. Vakali, and K. Kafetsios, "Micro-blogging Content Analysis via Emotionally-Driven Clustering," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013, pp. 375–380.

[56] J. Carrillo-de-Albornoz and L. Plaza, "An emotion-based model of negation, intensifiers, and modality for polarity and intensity classification," *J. Am. Soc. Inf. Sci. Technol.*, vol. 64, no. 8, pp. 1618–1633, Aug. 2013.

[57] C. Chiu, N.-H. Chiu, R.-J. Sung, and P.-Y. Hsieh, "Opinion mining of hotel customer-generated contents in Chinese weblogs," *Curr. Issues Tour.*, vol. 18, no. 5, pp. 477–495, May 2015.

[58] S. Schmunk, W. Höpken, M. Fuchs, and M. Lexhagen, "Sentiment Analysis: Extracting Decision-Relevant Knowledge from UGC," in *Information and Communication Technologies in Tourism 2014*, Cham: Springer International Publishing, 2013, pp. 253–265.

[59] A. Tripathy, A. Anand, and S. K. Rath, "Document-level sentiment classification using hybrid machine learning approach," *Knowl. Inf. Syst.*, vol. 53, no. 3, pp. 805–831, Dec. 2017.

[60] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley*

Interdiscip. Rev. Data Min. Knowl. Discov., Mar. 2018.

[61] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural Language Processing (Almost) from Scratch," *J. Mach. Learn. Res.*, vol. 12, pp. 2493–2537, 2011.

[62] W. Zhao *et al.*, "Weakly-Supervised Deep Embedding for Product Review Sentiment Analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 1, pp. 185–197, Jan. 2018.

[63] Y. Ma, H. Peng, and E. Cambria, "Targeted Aspect-Based Sentiment Analysis via Embedding Commonsense Knowledge into an Attentive LSTM," *AAAO-18*, 2018.

[64] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical Attention Networks for Document Classification," pp. 1480–1489, 2016.

[65] J. Xu, D. Chen, X. Qiu, and X. Huang, "Cached Long Short-Term Memory Neural Networks for Document-Level Sentiment Classification," Oct. 2016.

[66] Z. Li, Y. Zhang, Y. Wei, Y. Wu, and Q. Yang, "End-to-End Adversarial Memory Network for Cross-domain Sentiment Classification," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2017, pp. 2237–2243.

[67] S. Robertson, "Understanding Inverse Document Frequency: On theoretical arguments for IDF," *J. Doc.*, vol. 60, no. 5, pp. 503–520.

[68] F. Chaubard, R. Mundra, and R. Socher, "Lecture Notes: Part I," 2015.

[69] Q. Le and T. Mikolov, "Distributed Representations of Sentences and Documents."

[70] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of Tricks for Efficient Text Classification," 2016.

[71] V. Čekanavičius, "SEMINARO „LOGISTINĖ REGRESIJA SOCIALINIUOSE TYRIMUOSE" MEDŽIAGA," 2011.

[72] V. Čekanavičius, "LOGISTINĖ REGRESIJA SOCIALINIUOSE TYRIMUOSE."

[73] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," Dec. 2014.

[74] E. Cambria and A. Hussain, "Socio-Aff ective Computing 1 Sentic Computing A Common-Sense-Based Framework for Concept-Level Sentiment Analysis."

[75] "WHEN WORDS MATTER MOST: TAILORING DOMAIN-SPECIFIC DICTIONARIES WITH DECISION ANALYTICS Tailoring Domain-Specific Dictionaries with Decision Analytics," *Conf. Inf. Syst. Technol.*, 2015.

[76] "Amazon Reviews: Unlocked Mobile Phones | Kaggle." [Online]. Available: https://www.kaggle.com/PromptCloudHQ/amazon-reviews-unlocked-mobile-phones. [Accessed: 02-Mar-2018].

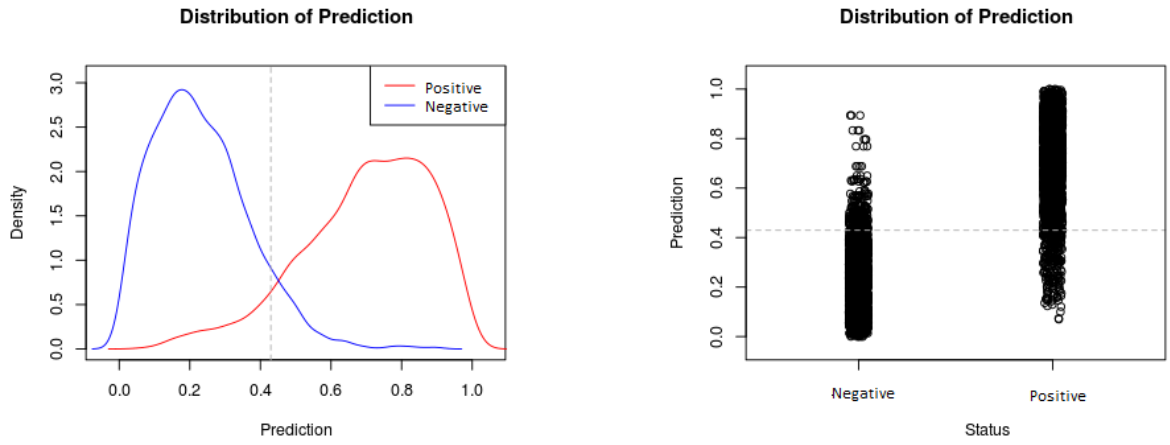
[77] "Bag of Words Meets Bags of Popcorn | Kaggle." [Online]. Available: https://www.kaggle.com/c/word2vec-nlp-tutorial/data. [Accessed: 02-Mar-2018].

[78] “TripAdvisor Reviews for The Eiffel Tower | Kaggle.” [Online]. Available: https://www.kaggle.com/PromptCloudHQ/tripadvisor-reviews-for-the-eiffel-tower/data. [Accessed: 02-Mar-2018].

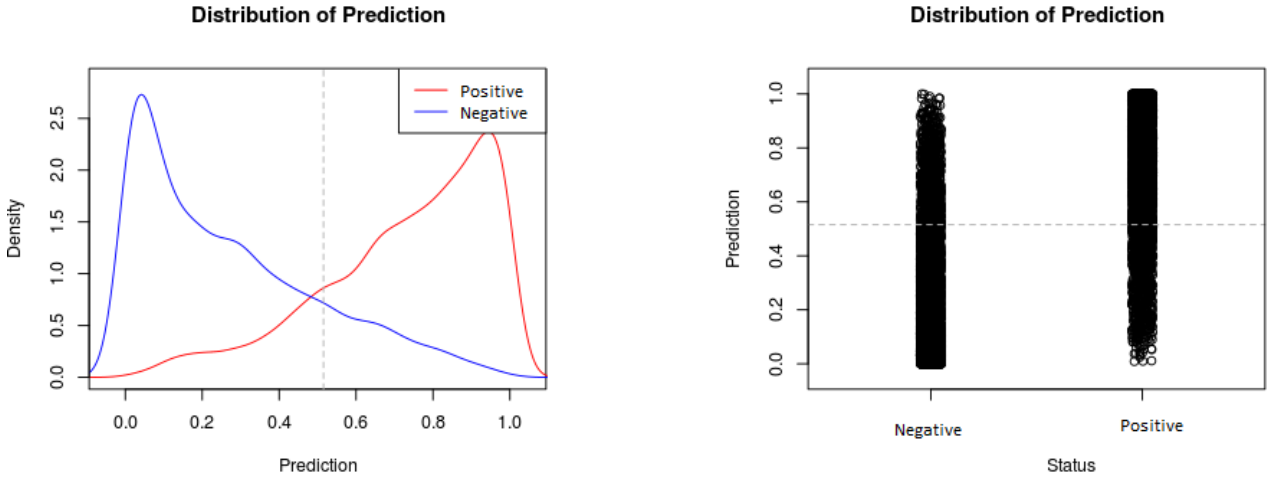
Priedai

1 priedas

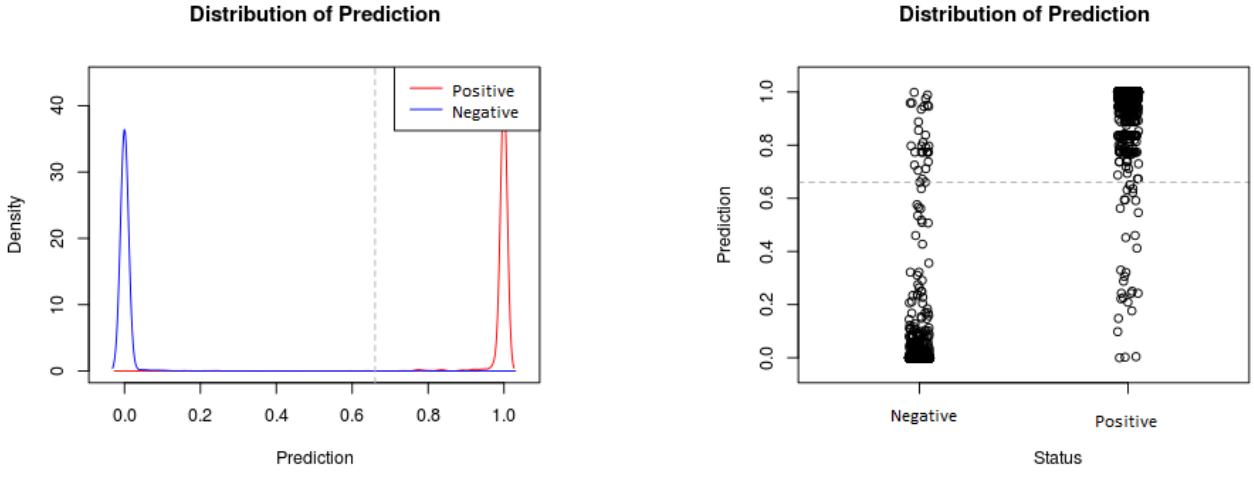
Klasių atskyrimo vertinimo diagramos



RF + RP metodų kombinacija su TripAdvisor duomenimis



RF + LSI metodų kombinacija su IMDB duomenimis



RF + Sen2Vec metodų kombinacija su Amazon duomenimis