Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Sentimento poliariškumo tyrimas Lietuvos įmonių klientų atsiliepimuose veidaknygėje ir evertink.lt

Baigiamasis magistro projektas

Laura Morkūnaitė

Projekto autorė

Doc. dr. Evaldas Vaičiukynas

Vadovas

Doc. dr. Aistė Dovalienė

Vadovė

Kaunas, 2019

Kauno technologijos universitetas
Matematikos ir gamtos mokslų fakultetas

Sentimento poliariškumo tyrimas Lietuvos įmonių klientų atsiliepimuose veidaknygėje ir evertink.lt

Baigiamasis magistro projektas
Didžiųjų verslo duomenų analitika (621G12002)

Laura Morkūnaitė
Projekto autorė

Doc. dr. Evaldas Vaičiukynas
Vadovas

Doc. dr. Aistė Dovalienė
Vadovė

Doc. dr. Kristina Šutienė
Recenzentė

Doc. dr. Beata Šeinauskienė
Recenzentė

Kaunas, 2019

Kauno technologijos universitetas
Matematikos ir gamtos mokslų fakultetas
Laura Morkūnaitė

Sentimento poliariškumo tyrimas Lietuvos įmonių klientų atsiliepimuose veidaknygėje ir evertink.lt

Akademinio sąžiningumo deklaracija

Patvirtinu, kad mano, Lauros Morkūnaitės, baigiamasis projektas tema „Sentimento poliariškumo tyrimas Lietuvos įmonių klientų atsiliepimuose veidaknygėje ir evertink.lt" yra parašytas visiškai savarankiškai ir visi pateikti duomenys ar tyrimų rezultatai yra teisingi ir gauti sąžiningai. Šiame darbe nei viena dalis nėra plagijuota nuo jokių spausdintinių ar internetinių šaltinių, visos kitų šaltinių tiesioginės ir netiesioginės citatos nurodytos literatūros nuorodose. Įstatymų nenumatytų piniginių sumų už šį darbą niekam nesu mokėjęs.

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

(vardą ir pavardę įrašyti ranka)	(parašas)

Morkūnaitė Laura. Sentimento poliariškumo tyrimas Lietuvos įmonių klientų atsiliepimuose veidaknygėje ir evertink.lt. Magistro baigiamasis projektas / vadovai doc. dr. Evaldas Vaičiukynas; doc. dr. Aistė Dovalienė; 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, vartotojų patirties valdymas, teksto vektorizavimas, mašininis mokymasis.

Kaunas, 2019. 63 p.

Santrauka

Klientų patirtis tapo viena iš pagrindinių tyrimų sričių, o patenkinti vartotojai įmonių siekiamybė. Sentimentų analizės pritaikymas versle gali būti įgyvendintas skirtingomis metodologijomis ir eigos etapais. Sentimento poliariškumo detekcijos uždavinys sprendžiamas mašininio mokymosi algoritmais ar leksikonu grįstais metodais.

Darbe bus tiriamas įvairių verslo sričių, internetinėje erdvėje palikti lietuvių kalba parašyti vartotojų atsiliepimai. Geriausios klasifikatoriaus kombinacijos sukūrimui buvo tikrinama hipotezė apie vektorizavimo dimensionalumo ir prognozavimo tikslumo priklausomybę. Išbandyti 8 skirtingi vektorizavimo metodai: žodžių krepšelio skirtingos modifikacijos, pastraipų vektorius – paskirstytos atminties metodas, latentinis semantinis indeksavimas, latentinis Dirichlė paskirstymas, atsitiktinių projekcijų metodas, Sent2Vec ir BERT. Sentimento poliariškumo detekcijos uždaviniui spręsti, klasifikatoriaus pagrindą sudarė atsitiktinių miškų, logistinės regresijos, atraminių vektorių ar gradientinio stiprinimo mašininio mokymosi algoritmai. Gautos kombinacijos buvo palyginamos tarpusavyje pagal Kappa tikslumo matą. Leksikonu grįstų metodų įgyvendinimui buvo pasirinkta generuoti specializuotą žodyną bei pasinaudoti tyrėjų pateiktais teigiamo ir neigiamo konteksto žodžių rinkiniais. Daugumoje atvejų, sentimento poliariškumą geriau identifikavo mašininio mokymosi ir žodžių išskyrimo į skirtingų dimensijų požymių vektorių, kombinacijos o ne žodynu grįsti metodai. Didžiausias tikslumas buvo pasiektas žodžių krepšelio metodo ir gradientinio stiprinimo modelio kombinacija.

Morkūnaitė, Laura. Sentiment analysis of Facebook and evertink.lt reviews for Lithuanian companies' customers. Master's Final Degree Project / supervisors: assoc. prof. Evaldas Vaičiukynas; assoc. prof. Aistė Dovalienė; 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, customer expierence, text data vectorization, machine learning.

Kaunas, 2019. 63 p.

Summary

Customer experience is one of the main research fields nowadays, and the aim of business is satisfied customers. Sentiment analysis may be implemented within different methodologies and contain various of stages of the process. Sentiment polarity detection can be solved by machine learning algorithms or lexicon-based methods.

Thesis work focus on e-reviews of Lithuanian various business fields companies' customers. For achieving the best classification combination, the hypothesis about accuracy and input data variables dimensionality correlation was tested. In the order, 8 vectors embedding methods were tried: Bag of Word method's modifications, Paragraph Vector Distributed Memory method, Latent Semantic Indexation, Latent Dirichlet allocation, Random Projections, Sent2Vec and BERT. The classifier's main component of the sentiment polarity detection task was Random Forest, Logistic Regression, linear Support-Vector machine or gradient boosting machine learning algorithms. All the possible classification combinations were compared within kappa accuracy score. Also, additional lexicon-based methods were utilized in this work. In the most cases, machine learning algorithm and document embedding combinations showed better results than lexicon-based ones. The most accurate results were achieved by word of the bag vectorization method and gradient boosting algorithm.

Turinys

Lentelių sąrašas

Paveikslų sąrašas

Santrumpos

SA – sentimento analizė;

DMPV – paskirstytos atminties pastraipos vektorius;

DBoW – paskirstytas žodžių krepšelis;

TF-IDF – termino dažnis – atvirkštinio dokumentų dažnis;

LSI – latentinis semantinis indeksavimas;

LDA – latentinis Dirichlė paskirstymas;

RP – atsitiktinis indeksavimas / atsitiktinės projekcijos;

LogR – logistinė regresija;

RF – atsitiktiniai miškai;

SVM – atraminių vektorių mašina;

XGBoost – gradientinis stiprinimas;

NB – Bajeso modelis;

PCA – pagrindinių komponenčių analizė;

API – aplikacijų programavimo sąsaja;

NLP – natūralios kalbos apdorojimas;

HARF – daugiabalsio nutarimo atsitiktinių miškų triukšmo valymo metodas;

Įvadas

Dinamiškai keičiantis vartotojų įpročiams, jų nuomonės sklaida persikėlė į virtualią erdvę. Žmonėms tapo paprasta, patrauklu ir lengva savo patirtas emocijas išreikšti e-kanalais, per verslo socialinių tinklų paskyras (pvz., *Facebook, Instagram*), forumus, asmeninius tinklaraščius. Vartotojas tokiu būdu yra įgalintas greitai, dažnai ir anonimiškai pateikti savo patirtį atsiradusią sąveikaujant su įmone fiziniuose ar virtualiuose kanaluose. Asmeninės nuomonės raiška, patarimų dalinimas, nuomonės formavimas tapo pagrindu susikurti tokio pobūdžio platformoms kaip forumai, diskusijų programėlės (pvz., *Discord*), o šių populiarumas patvirtina prielaida, kad žmonių tarpusavio bendravimas bei dalinimasis patirtimi yra svarbi dedamoji.

Vartotojai yra varomoji verslo jėga, be kurių tolimesnės verslo ateities perspektyvos būtų kvestionuotinos. Šiuolaikinės įmonės supranta ir vertina vartotojo pasitenkinimą, kadangi teigiamos vartotojo emocijos padeda įmonėms stiprinti jų lojalumą bei palaikyti darnius tarpusavio santykius. Sėkmingoms įmonėms privalu turėti klientų aptarnavimo, santykių palaikymo, CRM (*angl. customer relationship management*) sistemų vystymo strategijas. Pagrindinis visų įmonių tikslas bet kurioje pramonės šakoje yra suprasti savo vartotoją, jo individualumą ir tai pritaikyti gerinant kontaktą ir sąveiką su įmonės aplinka. Vartotojų santykių palaikymo esmė yra kurti, valdyti ir stiprinti lojalius bei ilgalaikius santykius. Tam, kad organizacijos gautų įžvalgų apie vartotojus atliekama vartotojų duomenų analizė, ne viena įmonė suprasdama naudą, įgalino nuomonės sklaidos galimybes – atsiliepimų skiltis. Nuomonės tyryba, kitaip vadinama sentimentų analize, padeda įmonei labiau įsigilinti į savo vartotoją, jį pažinti bei stiprinti ryšį pateikiant vartotojui būtent tai ko jis nori ar transliuojant tikslinei grupei pritaikytas žinutes. Tuo pačiu, tokių duomenų surinkimas leidžia identifikuoti ir neigiamas emocijas keliančius veiksnius. Sentimentų analizė yra viena labiausiai populiarėjančių, duomenų tyrybos CRM kontekste, sričių, o šios srities pritaikomumas marketinge yra gana naujas. Kompiuterių ir matematikos mokslas įgalina sukurti metodus, prognozavimo modelius ir kt., kurie padeda ištirti nuomonę ir gauti vertingų įžvalgų. Detekcijos uždavinio sprendimas rinkodaroje leidžia sukurti modelius, kurie geba atpažinti tikslinį objektą, pvz.: vartotojo nuomonės poliariškumą. Tinkamo klasifikatoriaus sukūrimas, leidžia pasiekti aukštus tikslumo rezultatus bei padaryti gilesnes įžvalgas, kurios įmonėms duoda pelno arba leidžia nepatirti nuostolių ir nenumatytų kaštų, o svarbiausia, užtikrina proaktyvius veiksmus.

Projekte naudojamas duomenų rinkinys sudarytas susisteminant skirtingų verslo sričių vartotojų atsiliepimus iš skirtingų duomenų šaltinių, kur 75 proc. rinkinio sudaro atsiliepimai surinkti iš evertink.lt nuomonės sklaidos tinklalapio ir 25 proc. iš socialinio tinklo „Facebook“ verslo paskyrų atsiliepimų skilties. Rinkinyje dominuoja tokie stambūs e-komercijos verslai kaip *pigu.lt* ar *knygos.lt*.

Darbo tikslas: sukurti tiksliausią atsiliepimo sentimento poliariškumo detektorių lietuvių kalbai.

Darbo uždaviniai:

1. apžvelgti literatūrą susijusią su sentimento analize ir jos taikymo metodais, ištirti jos naudą gerinant vartotojų patirtį.
2. išanalizuoti lietuviškų tekstų sentimentų analizės progresą Lietuvoje;
3. palyginti skirtingas vektorizavimo, dimensionalumo ir klasifikavimo modelių kombinacijas, randant tinkamiausias lietuviškiems tekstams;
4. aptarti surastas geriausias kombinacijas, įvertinant kalbinių modifikacijų perspektyvas;
5. pateikti siūlymus geriausios kombinacijos pritaikomumui verslo situacijose bei išskirti tolimesnes vystymo sritis;

1. Literatūros apžvalga

Kiekvienos įmonės tikslas yra sukaupti patirties ir ugdyti gebėjimus generuoti, skleisti ir panaudoti informaciją apie klientus, sukuriant jiems aukščiausią galimą vertę. Stiprioje konkurencinėje aplinkoje, įmonėms privalu nustatyti, kaip pagerinti vartotojų patirtį tam, kad būtų padidintas lojalumas, sumažinta klientų rotacija, sukuriama pridėtinė vertė. Pasinaudojus naujausiomis technologijomis, įdiegusiomis didžiųjų duomenų apdorojimo procesus ir prieiga prie klientų informacijos per įvairius Internetu grįstus kanalus, įskaitant e- ir m- komerciją, socialinius tinklus, daiktų internetą ir netgi lojalumo programas, atsiveria galimybės padaryti ankščiau neįmanomų įžvalgų apie klientus. Naujosios technologijos įgalina įmones rinkti ir analizuoti nefiltruotą vartotojų nuomonę, suprasti jų poziciją ir elgseną ir netgi užmegzti abipusį dialogą. Interneto, teksto, sentimentų, socialinių tinklų ir kt. tipo analitika yra naudojama analizuojant įvairių struktūrų klientų duomenis, tam, kad būtų sukuriami tobulesni prognozavimo modeliai, kas leidžia įmonei savo klientams pasiūlyti personalizuotus produktus ar paslaugas, proaktyviai numatyti ir atitikti klientų poreikius.

Technologijos leidžia automatizuoti įvairius procesus, nuo kliento duomenų rinkimo, valdymo, integravimo ir analizės iki tokios surinktos informacijos panaudojimo sprendimų priėmimo procesuose, kadangi patobulinimai leidžia tai vykdyti realiu laiku, pasinaudojant mašininio mokymosi algoritmais, kuriems dažniausiai nereikia net žmogiško ekspertinio vertinimo. Vartotojų analitika, paremta didžiaisiais duomenimis, ne tik pagerina įmonės veiklos rezultatus, tačiau ir kuria dar didesnį konkurencinį pranašumą, kurį sunku nukopijuoti.

1.1. Vartotojo patirties samprata

Per pastaruosius 25 m. sukaupta praktika ir atlikti moksliniai tyrimai, kardinaliai transformavo rinkodaros suvokimą, pakeitė stebėjimo rakursą ir padėjo sutelkti įmonių dėmesį į skirtingus objektus. Pradžioje, orientacija į produktą ir prekės ženklą, pakeitė stiprus dėmesys į santykių su vartotoju sukūrimą bei bendradarbystę, o šiuolaikiniai marketingo specialistai deda visas pastangas užtikrinti kuo geresnę patirtį vartotojui [1].

Klientų patirtis tapo viena iš pagrindinių tyrimų sričių, o patenkinti vartotojai įmonių siekiamybė. Tinkamos patirties vartotojui sukūrimas tapo nauja konkurencingumo erdve rinkodaroje, bei, tyrėjų nuomone, yra aukščiau už produkto ar paslaugos kokybę [1].

Tobulėjant technologijoms leidžiančioms įvertinti vartotojo būseną nesant žmogiškai interakcijai, vartotojų patirties suvokimas ir apibūdinimas tarp tyrėjų išsiskyrė. Abstraktų apibūdinimą pateikė Sharma‘s ir Chaubey [2] - tai visuma visų klientų ir prekių ir (arba) paslaugų teikėjų patirties, visu jų, kaip klientų, gyvavimo laikotarpiu. Tai gali apimti tokius emocinius ir fizinius pojūčius, kaip sąmoningumas, atradimas, patrauklumas, sąveikavimas, pirkimas, naudojimas, auginimas ir propagavimas. Kiek kitokį suvokimą galima rasti nagrinėjant XX a. pabaigos literatūrą. Vartotojo pasitenkinimas ir būsena ‘po’, buvo tirta ne vieno autoriaus, Vavra‘s [3] tuometinį fenomeną aprašo kaip 2 dedamųjų, - pasekmės ir proceso, - sąjunga:

- Pasekmės dedamoji tai kompleksinė įvairių vartotojo patirčių rinkinys. Rinkinį sudaro: kognityvinė vartotojo būklė įvertinus skirtas pastangas bei gautą vertę kontaktavimo su įmone periodu; emocinis atsakas į įgytą patirtį susietą su įsigytais produktais, paslaugomis, aplinka; pirkimo ir gautos naudos palyginimas, susidaręs iš vartotojo perspektyvos.

- Proceso dedamoji tai pasitenkinimas laikomas procesu, pabrėžiančiu suvokimo, vertinimo ir psichologinius procesus, kurie turi įtakos pasitenkinimui ir patyrimui.

Tiriant ir analizuojant vartotojų patirtį, svarbus aspektas suprasti savo įmonės klientą, personalizuoti juos, ištirti jų poreikius, pirkimo elgseną ir kt. Literatūroje galima rasti, kad neretai vartotojai yra kategorizuojami. Šiuolaikinėje rinkoje, taip pat kaip ir aprašoma teorijoje, įmonės stengiasi pažinti savo prekių ar paslaugų naudotoją, išskirti jo išskirtinius bruožus, juos segmentuoti, teikti prioritetą kuriai atitinkamai grupei. Identifikuoti vartotoją dažnai tampa sunkia užduotimi, nuo kurios sėkmės priklauso ir įmonės pelningumas ar netgi tolimesnis veiklos užtikrinimas. Vartotojai – įvairios socialinės klasės, išsilavinimo, tautybės ir kt. asmenys, kurie turi galimybę ir teisę priimti sprendimus. Tai grupė asmenų, kuriems yra poreikis patenkinti atsiradusius poreikius ar norus, o tai jie realizuoja pirkdami prekes, naudodamiesi paslaugomis. Paprastą, universalų vartotojų identifikatorių pateikė Siskos ir Grigoroudis [4] kurie susistemino kitų autorių darbus ir pasiūlė vartotojus išskirti į esamus, pasitraukusius ir potencialius:

- Esamais vartotojais laikomi tokie klientai, kurie per atitinkamą matavimo laiką, pirko ir naudojosi įmonės paslaugomis ar kitaip sąveikavo su įmone ir teikė pridėtinę vertę.

Numatytas laikotarpis svyruojantis ir priklauso nuo įmonės veiklos konteksto ar kitų kriterijų, pvz.: kavinių tinklas gali nustatyti ar klientas yra pastovus, nepasitraukęs, įvertindami ar jis apsilankė ir vartotojo įmonės produktus per pastaruosius 3 mėnesius. Ši vartotojų kategorija priskiriama prie vertingiausių, jiems turi būti skiriamas papildomas dėmesys, jų elgsena, pomėgiai, interesai yra aiškūs ir žinoma įmonei, ištirtas ir klientų atsakas į įvairias skatinimo priemones. Šių vartotojų išlaikymui gali būti skiriamas mažiau resursų, nei pritraukiant naujus klientus.

- Pasitraukiais vartotojais laikomi tokie klientai, kurie dėl įvairių priežasčių pasitraukė iš įmonės ir jos nebelaiko prioritetiniu pasirinkimu.

Ši grupė gana informatyvi, kadangi ištyrus jų pasitraukimo priežastis, jų patirtį visu jų gyvavimo periodu, leistų identifikuoti problemines įmonės sritis, pvz.: aptarnavimas. Yra tikimybė pastebėti, kad pardavimo procese yra kliūčių, kurios ir trukdo vartotojui priimti sprendimą pirkti iš įmonės.

- Potencialių vartotojų grupė, tai tokie asmenys, kurie pasiruošę pirkti siūlomą gaminį, tačiau jie vis dar renkasi arba laukia lūkesčius atitinkančio pasiūlymo.

Potencialių klientų paieška yra dinamiškas, ilgai trunkantis procesas, kadangi, visu pirma, esami klientai gali tapti buvusiais klientais, o norint sėkmingai vystytis, reikia turėti pirkėjų ar plėsti verslą pritraukianti vartotojus ir kitų rinkų. Ši klientų grupė taip pat įmonei suteikia informacijos, pvz.: faktorių, dėl kurių vartotojas nusprendžia pirkti, sužinojimas. Detalesnis vartotojų identifikavimu užsiima pačios įmonės, kadangi nuo veiklos srities vartotojų tipažas – skiriasi. Atlikti tyrimai norinti išsiaiškinti vartotojo segmento įtaką jo patirčiai: tyrimo objektu pasirinkta skirtingų veiklų [5], viešbučių tinklų [6], mobilios bankininkystės [7] vartotojai. Apibendrinus tyrimų rezultatus, gauta, kad segmentas/grupė turi reikšmingą įtaką ir vartotojų patirtis skiriasi tarp skirtingų tipų.

Vartotojai svarstydami, priimdami sprendimą naudotis įmonės paslaugomis ar pirkti prekes, patenka į įmonių sukurtą aplinką, kurioje jie siūlo savo paslaugas. Šioje aplinkoje vartotojai renkasi, įvertina, apgalvoja, tyrinėja, todėl patrauklios aplinkos sukūrimas ir vystymas yra svarbus faktorius, lemiantis vartotojo elgseną bei jo patirtį kontaktuojant su įmone. Aplinkos svarba yra neatmetama, kadangi

dažnai vartotojas ten patekęs, susidaro pirmąjį įspūdį, kuris gali būti lemtingas svertas jo apsisprendimui. Fizinė vartotojo patirtis siejama su kliento fiziologiniu atsaku į sąveika su aplinka [8]. Išskiriami aplinkos tipai:

- Fizinė aplinka - patalpos, įranga, darbuotojai

Fizinė aplinka gali būti ir virtuali, internetinės svetainės pavidalu. Technologijų plėtra įgalino įmones tiksliau ištirti aplinkos sąveiką su vartotojo pasirinkimu ir įtaką jo potyriams. Fizinės aplinkos svarba ypatingai pasijaučia sektoriuose, kuriuose svarbi estetika, dizainas ir kt., pavyzdžiui apgyvendinimo paslaugas teikiančiose įmonėse ar mados namuose. Neretu atveju, vartotojas įmonę renkasi būtent pagal jos fizinę aplinką (pvz.: viešbučio kambario dizainas), todėl nepaneigiamas faktas, kad norint pakelti vartotojų pasitenkinim1, būtina orientuotis į aplinkos tobulinimą. Kitas atvejis, kai įmonės fizinė aplinka yra virtuali. Šiais laikais, kada vartotojams nebeliko atstumo barjero, virtuali aplinka taip pat turi įtaką vartotojo pasirinkimui. Vartotojas rinksis jam patrauklesnį, patogesnį, lengvai suprantamo dizaino svetainę [9].

- Socialinė aplinka arba kitaip galima apibūdinti kaip kultūra įmonėje.

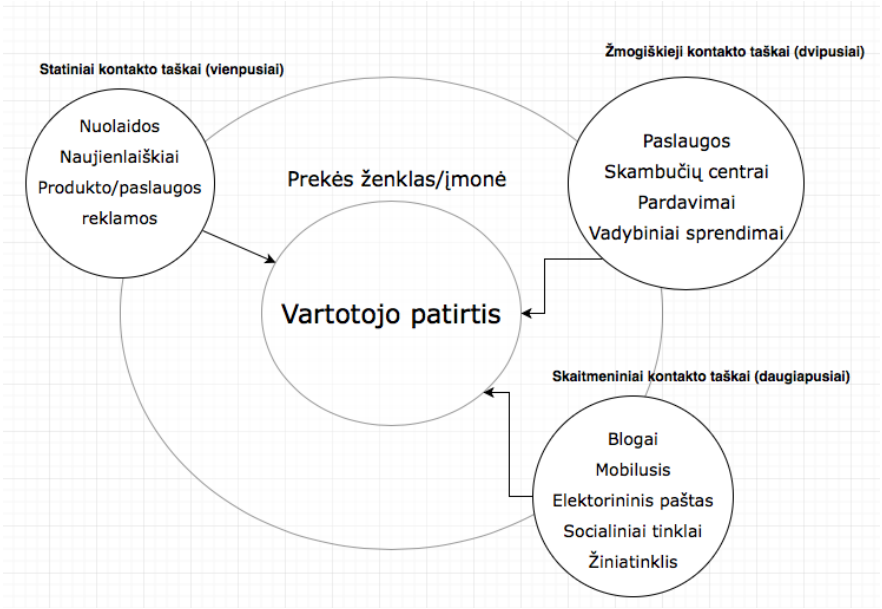
Šį tipą kurią, ugdo įmonės darbuotojai, jų elgesys, įmonės aptarnavimo standartai, abstrakčiai apibendrinus visas užmegztas ryšys su klientu. Šis socialinis komponentas grindžiamas santykių, atsirandančių tarp individo ir kitų žmonių, kurie bendrauja socialinėje aplinkoje, kokybe ir intensyvumu. Socialinė aplinka yra viena iš tų dedamųjų, kurią sunku nukopijuoti varžovams, nes šią aplinką sudaro ir visi užmegzti ryšiai su klientais, bei modelis, kaip priimtiniausia bendrauti su vartotojais. Įmonė sukūrusi patrauklę socialinę aplinką, ja gali naudotis kaip visapusišku pranašumu prieš kitus, kadangi tai nematomoji įmonės pusė. Santykių užmezgimas su vartotoju, įmonei duoda tik pranašumą, t.y. lojalius vartotojus, išaugusius pardavimus, kuriamas stiprus prekinis ženklas. Atitinkamai nagrinėjant mažmeninės prekybos rinką, papildomai įsitraukia ne tik kliento – darbuotojo, bet ir kliento-kliento socialinis ryšys. Šiuo ryšiu klientai užmezga ryšius dalindami/gaudami patarimus, išklausydami nuomonę [8].

Modernėjant žiniasklaidai, kai socialiniai tinklai tapo neatsiejama žmonių kasdienybė, įmonės turi galimybes pasiekti savo potencialių vartotojų ratą tiesiogiai ir greičiau, su didesniu jų atsaku. Tokios galimybių išnaudojimas, tiesiogiai siejasi ir su kuriama patirtimi. Vartotojų patirtis gali būti apibūdinama trimis skirtingomis fazėmis:

1. prieš-pirkimo;
2. pirkimo;
3. įsigijimo;

Pirmoji stadija tai prieš-pirkimo, kuri apibūdinimą kaip pirminis vartotojo susipažinimas su įmone, prekės ženklu. Pirminis kontaktas su vartotoju gali būti per visus ‘online‘ kanalus, pvz: reklamjuostes, nuomonės formuotojų sukurtus reklaminius pranešimus, reklama socialiniuose tinkluose, ar fiziniuose objektuose, tokiuose kaip reklaminis stendas autobusų stotyse. Tai pradžia formuojant vartotojo nuomonę ir įmonės įvaizdį. Pirkimas, tai kita stadija, kai yra tiriamas ir vertinamas vartotojo sąveikavimas su pačia įmone, jos prekiniu ženklu ir aplinka, kurioje vykdomi mainai. Tai yra tokios paslaugos, kaip pvz.: naršymas įmonės internetinėje svetainėje ar personalo aptarnavimas, kai yra vykdomas pirkimas. Po-pirkimo stadija apibūdinama kaip kliento nuomonės išreiškimas, iškeltų

pradinių hipotezių patvirtinimas arba paneigimas ir visa kita likusi sąsaja su įmone. Kelionės metu nuo išankstinio pirkimo iki įsigijimo prekės ženklai turi įvairias galimybes pagerinti savo klientų patirtį. Vartotojai yra linkę savo patirtimi pasidalinti, ypatingai neigiama, socialiniuose tinkluose, forumuose ar kitaip kontaktuojant su kitais suinteresuotais asmenimis [10]. Visu kliento kelionės periodu, jis kontaktuoja su įmone per įvairius kontakto taškus. Vartotojo kontakto taškas, tai įmonės galimybės surasti ir sukontaktuoti su potencialiu ar esamu vartotoju. Kiekvieną kartą, kai klientai pamato prekinį ženklą internete ar skelbimuose, peržiūri reitingus ir apžvalgas, apsilanko internetinėje svetainėje, apsiperka parduotuvėje ar komunikuoja kliento aptarnavimo klausimais, galima teigti, kad klientas naudojasi įmonės kontaktiniais taškais. Tokių taškų sąrašas gali būti ilgas ir priklauso nuo pačios įmonės sugebėjimo rasti sprendimus kur realizuotis. Pateikiamos apibendrintos kontakto taškų grupės bei jų sąsaja su vartotojų patirtimi (pav. 1) [11].



1 pav. Kontakto taškų ir vartotojo patirties sąveika

Kaip ir minėta ankščiau, socialiniai tinklai yra vartotojų kasdienybė. Įmonių tikslas palaikyti darnius, informatyvius santykius šiame kanale, kuriant išskirtinius santykius visu vartotojo kelionės metu.

Detaliau nagrinėjant šį fenomeną, autoriai pritaria, kad interneto ir socialinės žiniasklaidos naudojimo augimas suteikia naujų galimybių. Skirtingai nei tradicinėse žiniasklaidos priemonėse, klientų ir firmų sąveika socialinės žiniasklaidos pagalba yra abipusiai naudingi mainai. Pirmiausia, komunikavimas su vartotojais jiems leidžia duoti atsaką į įmonės produktus, kainą ir kt., toliau socialiniai tinklų ir nuomonės formuotojų pagalbą, stiprinamas prekės ženklo įvaizdis. Tokie asmenys, kurie formuoja ir turi įtakos tikslinės grupės nuomonei, yra priskiriami prie įmonės advokatų segmento. Galiausiai, įmonės turinys nagrinėjamoje platformoje ir klientų atsakas, padeda įmonei suprasti ir atnaujinti informaciją apie savo tikslinės auditorijos pomėgius, elgseną ir kt. [12].

Platformose pastebimas klientų paliekamų atsiliepimų padidėjimas, jie linkę skleisti ir aprašyti savo patirtį. Dėl šios priežasties, rinkodaros specialistams - tai naujų tyrimų erdvė. Šalyse socialinės žiniasklaidos platformų populiarumas skiriasi pagal platformas. Internetinės svetainės "Fortune 500“[1] duomenimis, iš visų šiame sąraše atsidūrusių įmonių, daugiau nei 77% aktyviai naudojasi „Twitter“

[1] Didžiausią pelną generuojančių JAV įmonių statistikos tinklalapis: http://fortune.com/fortune500/

socialine platforma, bei ši platforma yra populiariausia tarp visų pasaulio socialinių tinklų, tiriant tik sąraše esančias įmones [13].

Apibendrinant, teigiama vartotojų patirtis yra konkurencinis įmonės pranašumas. Įmonės galinčios užtikrinti sklandų bendradarbiavimą su klientais, pasiūlydamos rinką ir poreikius atankančias prekes, turi didesnę tikimybę sėkmingiau vystyti savo verslą, būti pelningais. Svarbu įvertinti, kad patirties kūrimas yra dinamiškas procesas, apimantis skirtingus kanalus, per kuriuos galima pasiekti vartotoją. Modernėjant ir tobulėjant procesams, technologijoms, keičiantis estetikos, dizaino suvokimui, būtina proaktyviai reaguoti į visus įmanomus pasikeitimas, bei užtikrinti tendencingą prisitaikymą. Technologijų dėka, vartotojo elgseną, pomėgius ir kt. tampa vis lengviau ištirti bei tyrimo rezultatus pritaikyti vystant verslą ir užtikrinant kokybiškus santykius su vartotoju.

1.2. Sentimentų analizė vartotojų patirties identifikavimui

Daugėja tyrimų, kuriuose identifikuoajama internetinėje erdvėje skleidžiamos ‘iš lupų į lūpas‘ (angl. *word-of-mouth*) nuomonės svarba verslui. Vienas iš tokios sklaidos būdų – vartotojų atsiliepimai. Atsiliepimas, tai pozityvų ar negatyvų kontekstą apie produktą, įmonę turintis objektas, kurio autorius yra potencialus, esamas ar buvęs klientas. Tokioje nuomonės raiškos formoje, vartotojas išreiškia savo po-pirkimo būseną bei pateikia savo vertinimą, kas yra apibrėžiama kaip jo patirtis. Dažniausiai atsiliepimas yra pateikiamas reitingavimo arba laisvo teksto formoje. Stebėti vartotojų paliekamus atsiliepimus itin svarbu verslui, kadangi vartotojas savo atsiliepimu perduoda informaciją apie verslo objektus, pvz., prekių kokybę ar aptarnavimą. Tuo pačiu, atsiliepimai formuoja įmonės reputaciją, o masiškas neigiamos nuomonės skleidimas gali privesti įmonę prie bankroto [14]. Atsiliepimas yra vertinamas kaip vartotojo patirties išraiška ir norint pagerinti apsipirkimo ir santykių palaikymo kokybę tarp kliento ir įmonės, dėmesys turi būti skiriamas išsiaiškinti veiksnius, kurie sukėlė būtent tokią vartotojo patirtį ir tokiais sentimentais išreikštą atsaką. Rinkodaroje jau seniai tiriama ir ieškoma inovatyvių ir tikslesnių būdų identifikuoti vartotojo sentimentą ir išgeneruoti jį paveikusius veiksnius. Patirtis rodo, kad vartotojai labiau linkę skleisti neigiamą emociją, tokia informacijos sklaida sukelia santykinai didesnę žalą nei teigiamų emocijų sklaida duoda naudos. Ištyrusios vartotojo neigiamą atsaką, įmonės gali pastebėti problemines savo verslo sritis bei jas koreguoti, proaktyviai reaguoti į pokyčius, identifikuoti sentimento poliariškumą lemiančius veiksnius, tobulinti bei vystyti į vartotojui palankesnę poziciją.

Tyrėjai Mäntylä, Graziotin ir Kuutila [13] išsamiai nagrinėjo literatūrą ir sentimento analizės panaudojimo dinamiką. Autoriai tyrinėdami paieškos platformas, mokslius straipsnius, pastebėjo, kad pastaruosius 20 m. išsaugo informacijos ir panaudojimo atvejų susijusių su sentimentų tyrybą. Jie sugebėjo rasti daugiau nei 7 tūkst. straipsnių šia tema, bei pastebėjo, kad 99 proc. buvo išleisti 2004 m. arba vėliau. Sentimentų analizė tapo viena populiariausių ir augančių tyrimo sričių. Sentimento analizė, tai subjektyvios nuomonės tyrimo analizė, kurioje pagrindinis dėmesys skiriamas teigiamų ir neigiamų nuomonių, emocijų ir vertinimų, išreikštų natūralia kalba, teksto pavidalu, nustatymui. Tai buvo vienas pagrindinių metodų, taikant ir atpažįstant tekstą pranešimuose, žinutėse, nustatant internetinių diskusijų sentimentalumą, klasifikuojant teigiamus ir neigiamus atsiliepimus. Dauguma šios analizės darbų ir tyrimų yra susijusių su sentimento nustatymu dokumentuose, pvz.: atsiliepime, tačiau tokie analizės pritaikymai tiriant, pvz., sentimentą, kai yra atsakinėjami klausimynai (angl. *Opinion Question Answering*) ar peržiūrėtos teksto tyrybos taisyklės, kai ieškoma nuomonės apie produktą ar įmonę, reikalauja analizės sakinio arba frazės lygiu [15].

Nagrinėjant darbus susijusius su sentimentų analize, randama su analize susijusių sąvokų.

- Sentimentas – išreikšta emocija, nuotaika, požiūris, jausmas žodine, rašytine ar kita priimtina forma, kurios tikslas perteikti savo patirtį. Sentimentas dažnai yra nukreiptas į konkretų objektą ir yra vertingas [16].

Natūralios kalbos apdorojimas yra sudėtinga procedūra, todėl ne visada lengvai išskiriamos sentimento dedamosios. Sentimentą galima išreikšti ir penkių komponenčių struktūra [17], kur e – sentimento objektas (targetas), a – objekto savybė, veikianti sentimento atsiradimą, s – išreikštas sentimento kontekstas apie objektą, h –sentimento skleidėjas (asmuo, pateikęs/išreiškęs sentimentą) ir t – laikas, kada buvo išreikštas sentimentas.

Šios formos naudojimui būtina suprasti pagrindinius principus: visos komponentės turi sąveikauti viena su kita bei jas sieja nenutraukiamas ryšys. Visų penkių komponenčių ištyrimas ir radimas sentimente yra svarbus uždavinys, kadangi vienos iš jų nebuvo atveju yra prarandama esminė ir informatyvi informacija, pvz.: nesant laiko komponentei, neįmanoma identifikuoti kitimo chronologijos, prarandama galimybė susieti sentimentą su aplinkos veiksniais, bei įvertinti sentimento reikšmingumą. Atveju, kai nėra objekto komponentės, sentimentas iš dalies praranda esminę naudą ir tampa abstraktus, kadangi tampa ypač sunku surasti ryšį. Praktikoje, norint išskleisti sentimentą į jo pateiktą struktūrą, gaunamas sudėtingas uždavinys dėl kalbos subtilybių, daugiaprasmybių, bei nežinomų aplinkybių.

Nuomonės sklaidai persikėlus į skaitmeninę, internetinę erdvę ir tiriant socialinius žiniatinklius, sentimento struktūros žymėjime išsiskiria komponentės, iš ankščiau pateikto komponenčių rinkinio [16], šio mokslininko pateiktą rinkinį sudaro: duomenų šaltinis (sakinys, dokumentas; forumas, socialinis tinklas), sentimento skleidėjas, sentimento objektas, išreikštos emocijos tipas(ai) (mėgsta, nekenčia, dievina ir kt.) ir poliariškumas.

Neretai sentimentą galima vertinti ir per abiejų komponenčių rinkinių prizmę. Sentimento intensyvumas nustatomas įvertinus ir ištyrus sentimento poliariškumą.

- Sentimento poliariškumas - nuomonės orientacija (teigiama / neutrali / neigiama) išreikšta natūralia kalba.

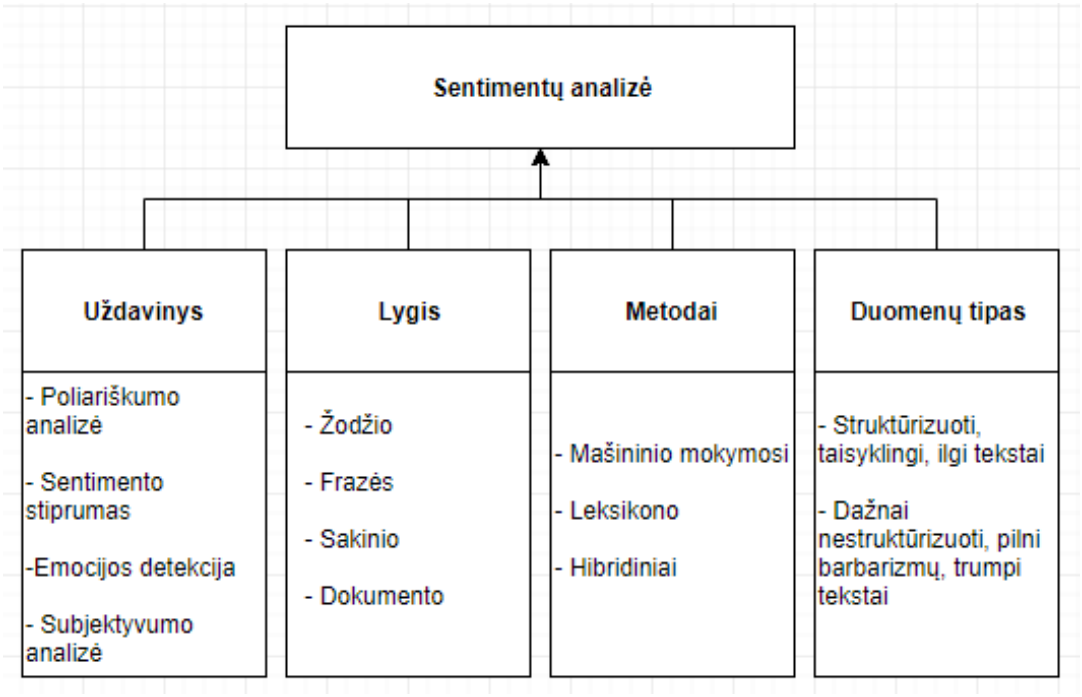
Sentimento poliariškumas daugeliu atvejų gali būti klasifikuojamas į teigiamą, neutralų ir neigiamą. Atitinkamai nusprendus kokie metodai taikomi ir kokį uždavinį norima spręsti, pvz.: detekcijos uždavinio atveju, sentimento poliariškumas gali būti vertinamas 2 klasėmis – teigiama ir neigiama.

Poliariškumą lemiančių veiksnių tyrimas yra dedamoji sentimentų analizėje. Tirti poliariškumą frazės lygiu yra kompleksinė ir sunki užduotis, kurioje gali pasitaikyti daug dviprasmiškumo. Sunku apibūdinti visus veiksnius, kurie galėtų veikti neigiamą poliariškumą, dažnai tiriant yra sunku tiesiogiai pastebėti neigiamą ar teigiamą kontekstą. Negatyvas gali būti identifikuojamas skirtingai, vienu atveju tai pastebima lokaliai, pavyzdžiui, išreikšta vienu žodžių – ‘negerai‘, ar ilgesniu, bendresniu teiginiu – ‘neatrodo gerai‘, ar tai galima pastebėti sakinyje ‘niekas negalvoja, kad tai yra gerai‘. Tačiau yra frazių, kuriose apstu neigiamo konteksto žodžių, tačiau pati frazė yra teigiama, pvz.: ‘tai negerai, o tai tiesiog nuostabu‘, o sarkazmo atveju – atvirkščiai, pvz.,‘labai geras, tik iškart sugedo‘. Konteksto poliariškumą, taip pat gali veikti modalumas, kai esmė yra išreikšta veiksmažodžio nuosakomis, dalelytėmis, intonacija, loginiu kirčiu ir kt. Be modalumo, kalboje susiduriama su žodžių prasme, žodžio sintaksiniu vaidmeniu sakinyje, pvz.: Eglė yra vardas ir tuo

pačiu medžio rūšis, todėl apie kurį objektą kalbama, tampa sunku identifikuoti, ypač kai jis vartojamas sakinio pradžioje arba blogas (reiškiantis neigiamą emocija, apibūdinimą) ir blogas (reiškiantis asmeninį internetinį dienoraštį). Tuo pačiu, poliariškumą gali veikti ir tematika – egzistuoja žodžių, kurių reikšmė sakinyje gali būti teigiama arba neigiama, kalbant apie skirtingas veiklos zonas. Reikėtų nepamiršti įvertinti ir asmens, kuris išreiškia sentimentą, poziciją. Pavyzdžiui, kalbant apie šių dienų aktualias naujienas, tokias kaip „Brexit“ bei Jungtinės Karalystės parlamento atmestą pasiūlytą pasitraukimo sutartį, parlamento nariams, kurie balsavo prieš šią sutartį, tai yra teigiamo konteksto naujiena, o JK ministrei pirmininkei – neigiamo. Todėl šiuo atveju, galima teigti, kad pats poliariškumas priklauso dar ir nuo to, kas išreiškia savo poziciją [15].

1.3. Sentimentų analizės taikymo kryptys

Pats sentimentų analizės apibūdinimas gali priminti apie esminius sentimentų analizės uždavinius, tai nuomonės tyrybą įvertinant emocijos tipą. Sentimentų analizė reikalauja daug informacijos, kuri yra gaunama iš skirtingų šaltinių ir apie skirtingas temas, todėl reikia visą informaciją susisteminti. Kiekvienais metais publikuojama daug straipsnių, kuriuose apžvelgiamos įvairių aspektų ir dimensijų problemos. Identifikuoti keliamos problemos, kurias norima išspręsti, tipą, yra vienas iš pagrindinių žingsnių prieš pradedant taikyti analizės pritaikymui reikalingus metodus. Pati sentimentų analizė, gali būti išskirstyta į pagrindines dedamąsias, pateikiamas ryšių grafikas, kuris atvaizduoja užduoties kryptingumo galimybes (2 pav.):



2 pav. Pagrindinės sentimentų analizės problemų kryptys (sudaryta autorės pagal [21])

Iš pradžių būtina kritiškai įvertinti, kurią užduoties kryptį nori ištirti. Išskirtos vienos populiariausių krypčių (angl. *subtask*), tačiau neretais atvejais, kryptingumas buvo kompleksinis, t. y., vienu tyrimu norima nustatyti ir poliariškumą, ir sentimento stiprumą. Poliariškumo analizės metu, pagrindinis uždavinys nustatyti ar sentimentas teigiamas, ar neigiamas. Sentimento stiprumo nustatymo būdu norima įvertinti, kiek stipriai sentimentas kaupia savyje nustatytą poliariškumą, kitaip tariant, atsako į klausimą: kaip stipriai yra teigiamas ar kaip stipriai yra neigiamas. Emocijos detekcija leidžia identifikuoti, kokia emocija ar jausmai buvo rašant ar kitokia forma pateikiant sentimentą (liūdesys,

laimė ir kt.) Subjektyvumo analizės krypties tyrimais norima įvertinti ar nagrinėjamas tektas subjektyvus (turi pozityvų ar negatyvų sentimentą), ar objektyvus (turi neutralų sentimentą) [21].

Aptariant kitą dimensiją, t. y. lygio dimensija, nustatoma, kokia yra taikymo sritis, identifikuojama tiriama objekto klasė / lygis. Mokslininkai Sun‘as, Lu (Luo) ir Chen‘as [22] ištyrę žinomas natūralios kalbos apdorojimo technikas, pateikė dokumento ir sakinio lygių apžvalgą. Dokumento lygiu sentimento analizės atliekamos pagal išgaunamą poliariškumą visame dokumento kontekste, o ne tik ieškoma neigiamą prasmę turinčių žodžių (pvz.: žodžio lygiu). Dokumento pavyzdžiu galėtų būti filmo apžvalga, atsiliepimas apie prekes, socialinio tinklo „Twitter“ trumpieji pranešimai ar blogo įrašai. Remiantis Liu [17] pateiktu nuomonės tyrybos penkių komponenčių rinkinio apibrėžimu, tiriant dokumento lygiu, būtent identifikuojama trečioji, išreikšto sentimento konteksto apie objektą, komponentė. Mokslininkai vienareikšmiškai pritaria, kad dokumento lygiu sentimento analizės tikslumo matais gauti rezultatai, dažnu atveju būna žemesni nei tiriant, pvz., tik sakinio lygiu. Sentimento poliškumas gaunamas suvidurkinus visų dokumente esančių sakinių nustatytą nuomonės orientaciją [21]. Analizės taikymas sakinio lygiu turi daug panašumų su dokumento lygiu, kadangi sakinio lygį galima traktuoti kaip mažą dokumentą. Klasifikavimas į dokumento ar sakinio lygius:

- dokumento lygis nustatomas, kai sakinių > 1.
- sakinio lygis nustatomas, kai sakinių = 1 ir frazių > 1.

Žodžio lygiu ištirti poliariškumą viena lengvesnių užduočių, kadangi yra paruošti žodynai, kuriuose pateikti visi teigiami ir neigiami žodžiai, pvz., Chen‘as ir Skiena [23] nepatingėjo tą atlikti visom didesnėm kalbom, įskaitant ir lietuvių.

Toliau [21] nagrinėja metodus, kuriais vykdoma sentimentų analizė. Nors matematinių algoritmų taikymo bumas siejamas su *Web 2.0* pradžia, tačiau didžioji dalis jų buvo atrasti ir mokslininkų aprašyti dar ankščiau nei 2000 m. [24].

Metodai suklasifikuojami į mašininio mokymosi, t.y. didžioji dalis algoritmų, kurie apibūdinami kaip ‘su mokytoju‘ arba kitur galima rasti ‘prižiūrimi“. Mašininio mokymosi algoritmo modeliai, veikiami klasifikavimo uždavinio mokomi ant pateikiamos duomenų imties, kai įvyksta apmokymo faktas, toliau pateikiama nematyta duomenų imtis, kurią išmokytas mašininio mokymosi modelis, geba suklasifikuoti pagal anksčiau pasiektą tikslumą, su sąlyga, kad modelis nepersimokė, nes tokiu atveju testavimo tikslumas stipriai krenta. Pagrindinis leksikonų metodų principas, kad nustatyta žodžių ar sakinio nuomonės orientacija, suvidurkinama ir priskiriama visam dokumentu

Pastebima, kad šis metodas nėra vienas tiksliausių, lengva suklysti dėl kalbos subtilybių, minėtų 1.2 poskyryje. Detaliau apie taikomus metodus pateikiama 1.4 poskyryje.

Galiausiai nagrinėjami duomenų tipai, duomenų įvestis. Dažnai praktikoje norint taikyti įvarius metodus, sprendžiant įvairių verslo sričių uždavinius, susiduriama su duomenų kokybės problema. Verslas duomenis kaupia neatsakingai, dažnai nestruktūrizuotai, dideli masyvai informacijos praranda ryšius struktūrinėse duomenų bazės. Sentimentų analizėje taip pat susiduriama su įvairios kokybės duomenimis. Duomenų įvairovė išaugo dar labiau, prasidėjus internetiniam amžiui. Sentimentų analizėje naudojami atsiliepimų, forumų, diskusijų, naujienų portalų komentarų duomenys gali būti suklasifikuoti į 2 plačias kategorijas, pirmoje pateikiami ilgi tekstai, kurie dažnai būna parašyti taisyklinga kalba bei paisoma sakinio struktūros reikalavimų. Antroje kategorijoje

atsiduria trumpi („Twitter“ žinutės apribojamos iki 140 simbolių), pilni barbarizmų, jaustukų (pvz. ‘:D‘, ‘xD‘ ir kt.), neaiškios, nevartotinos kalbėsenos tekstai.

Apibendrinus, sentimentų analizės identifikavimas leidžia rasti sąsajų su ankstesniais kitų autorių tyrimais, juos palyginti tarpusavyje, įverti tyrimų rezultatus bei suprasti kas geriausiai tinka nustatytai problemai spręsti.

1.4. Sentimentų analizės eigos ir metodai

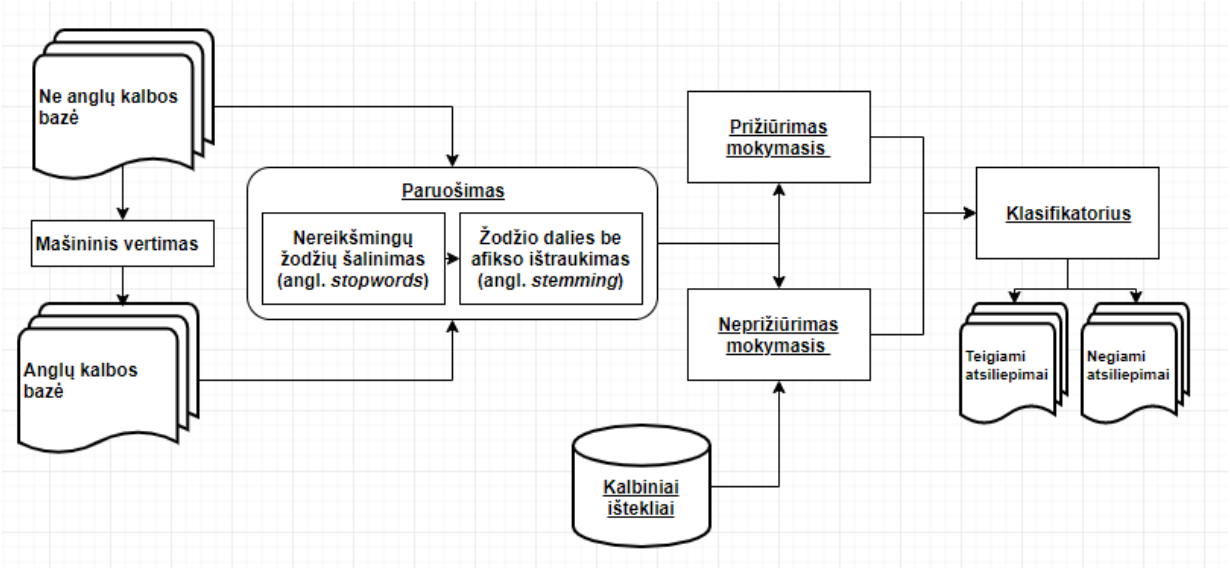
Sentimentų analizė automatiškai klasifikuoja dokumente išreikštas nuomones, dažniausiai į teigiamas arba neigiamas klases. Dokumentai, kurie yra vartotojų atsiliepimai, apskritai atspindi kliento nuomonę apie tekste paminėtus objektus. Todėl automatizavus sentimentų analizės vykdymą, galima turėti naudingą įrankį, kuris pvz.: leistų atlikti greitą paiešką internete pagal nuomonę arba automatinį atsiliepimų suklasifikavimą ir atskyrimą į norimas klases. Nors sentimentų analizė yra labiau teksto klasifikavimo uždavinys, tačiau kaip jau minėta ankščiau, tekstinis formatas būna įvairus. Tam, kad kokybiškai atlikti, sentimentų analizės uždaviniui būtina pritaikyti skirtingo tipo metodus, kadangi metodų veikimo principai skirtingai reaguoja į duomenis reprezentuojančius parametrus [25].

Siekiant nustatyti konkrečios nuomonės orientacija, išskiriamos 2 metodų kryptys, kaip jau buvo aptarta 1.3 skyriuje, tai mašininio mokymosi ir leksikono metodai. Toliau nagrinėjant mašininio mokymosi algoritmus, juos gali išskirti į 3 segmentus:

1. prižiūrimas mokymasis;
2. pusiau prižiūrimas mokymasis;
3. neprižiūrimas mokymasis;

Prižiūrimo mokymosi atveju, dažnai turimas klasifikavimo uždavinys, kuris įprastai perklasifikuojamas į dektekcijos uždavinį. Sentimento stiprumui įvertinti, turimas regresijos uždavinys [21].

Potencialią proceso eigą ir galimas pasirinkimų opcijas susistemino ir pateikė Martín-Valdivia, Martínez-Cámara, Ortega ir López [26] savo darbe, kuriame nagrinėjo poliariškumo detekcijos problemą bei palygino ir apjungė prižiūrimo ir neprižiūrimo mokymo algoritmus, įvertinant atsiliepimų sentimento orientaciją ispanų kalboje.



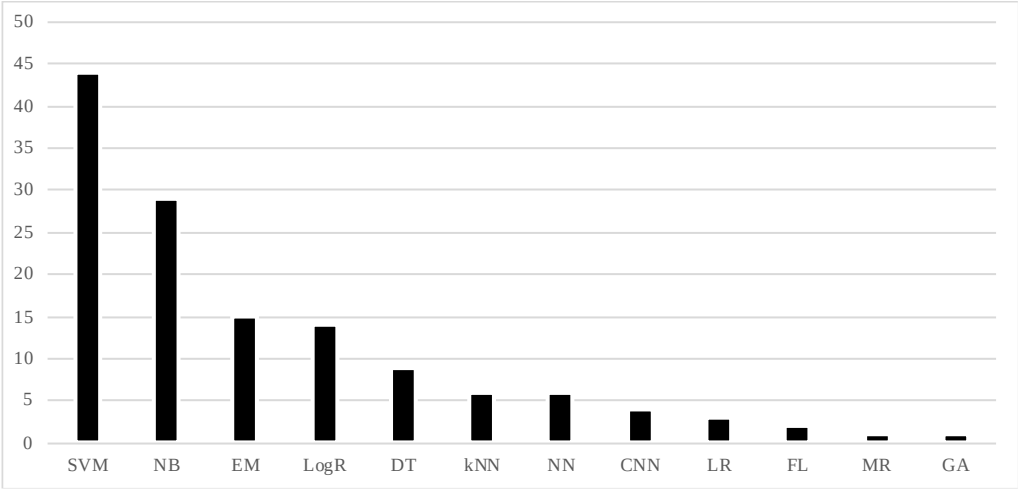
3 pav. Sentimento analizės eiga (sudaryta autorės pagal [26])

Pasirinktų duomenų kalbinė bazė svarbus aspektas taikant sentimentų analizę. Masiškai didėjant socialinių tinklų vartotojams, ekonomiškai stipriose, augančiose šalyse, dauguma gyventojų turi gebėjimų bendrauti daugiau nei viena kalba, kitaip įvardinant yra dvikalbiai ar daugiakalbiai. Ypač socialiniuose tinklų duomenyse, pastebima tendencija mintis, emocijas, patirtį reikšti keliomis kalbomis. To pasekmė yra daug kompleksiškesnė sentimentų analizė, kai sugeneruojami duomenys dažnai turi skirtingas gramatines taisykles ir kt., o tai tampa papildoma užduotimi apdorojant natūralią kalbą [27]. Mokslininkai Martín-Valdivia ir kt. [26] siūlo ne anglų kalbos bazę papildomai išsiversti automatinių, mašininių vertėjų pagalba. Tokiu atveju, sukuriamas vertimo mechanizmas (pvz.: pasinaudojus Python gilaus mokymosi biblioteka Keras sukuriamas transformeris) arba pasinaudojus jau sukurtais mechanizmais (pvz., Google vertėjas).

Toliau, priklausomai nuo tyrėjo pasirinkimų, sentimentų analizę galima išskirti į prižiūrimą ir neprižiūrimą. Sentimento analizė išsiskaido į dvi skirtingas metodologijų kryptis, tai leksikonu grįstus metodus, kuriuos dauguma atvejų galima klasifikuoti kaip neprižiūrimus, ir prižiūrimo mašininio mokymosi algoritmus, kitaip vadinamus - ‘su mokytoju‘. Nustatyti sentimento poliariškumą gali būti sunki užduotis netgi ir žmogui. Prižiūrimo sentimento analizės pagrindinė kryptis - mašininio mokymosi algoritmais sukurti modelius su anotuotais duomenimis. Duomenys yra naudojami apmokymo etapui, siekiant įvertinti jų sintaksinę struktūrą, naudojant duomenyse esančius žodžius ar sakinius kaip kintamuosius, prieš tai juos papildomai apdorojant, pvz., vektorizuojant į nustatytą dimensionalumą. Pagrindinis tikslas, sukurti modelį, kuris gebėtų su dideliu tikslumu nustatyti dokumento poliariškumą. Šiai procedūrai atlikti yra naudojami klasifikavimo algoritmai, o šios procedūros integravimas į realius kanalus, leidžia greitai analizuoti ir ištirti dinamiškoje aplinkoje esančius dokumentus, tekstus, atsiliepimus. Šio metodo pagalba galima tiksliau identifikuoti sentimentą ir neretai išvengti įvairių kalbos subtilybių, pvz.: sarkazmo tekste, bei tokį atsiliepimą tiksliai klasifikuoti, kas nepavyktų su leksikonu grįstais metodais. Sukurti modeliai taip pat gali būti kiekvieną kartą modifikuoti, juos papildomai apmokant vis naujais, dar nerodytais duomenimis, taip stipriai pagerinant modelio prognozavimo / klasifikavimo tikslumą [28].

Prižiūrimai sentimentų analizei pritaikyti ir atlikti galima naudotis skirtingus klasifikavimo algoritmus. Nagrinėjant literatūrą ir mokslinius darbus apie sentimentų analizę, dažniausiai taikomi prižiūrimo mokymosi algoritmai: statistiniai (regresija, Bajeso, SVM ir kt.), struktūriniai (taisyklėmis

grįsti, atstumu grįsti ir kt.) ir modeliai – ansambliai (bagging, boosting, atsitiktiniai miškai). Analizuojant „Twitter“ platformos duomenis, geriausi rezultatai buvo pasiekiami su atraminių vektorių mašina (angl. *Support Vector Mashine*), toliau SVM, taip pat pakankamai tikslūs rezultatai buvo pasiekiami Bajeso (angl. *Naïve Bayes)*, toliau NB, ir modeliais – ansambliais (EM) [29]. Šie autoriai savo darbe susistemino pastarojo dešimtmečio mokslinę literatūros apie sentimentų analizę, bei išskyrė dažniausiai naudotus algoritmus (4 pav.).



4 pav. Dažniausiai naudoti prižiūrimo mašininio mokymosi modeliai sentimentų analizėje (sudaryta autorės pagal [29])

Iš paveikslo matoma, kad populiarūs yra ir artimiausių kaimynų (KNN), ir neuroninių tinklų (NN, cNN) algoritmai. Dažnai, papildomai norint optimizuoti modelių parametrus, taikomi genetiniai algoritmai (GA). Genetinių algoritmų esminis principas, tai gamtoje sutinkami reiškiniai, kurių elgesys paremtas kai kurių gamtos evoliucijos mechanizmų metafora. Paprastesnis optimizavo pritaikymas gali būti įvykdytas su atsitiktine (*angl. random*) arba tinklelio (*angl. grid*) paieška.

Toliau, statistiniai metodai SVM ir NB, buvo pritaikyti ir analizuojant bei bandant nustatyti sentimentus komentaruose, parašytuose kinų kalba. Gauti n-gramų junginiai, kurie leido identifikuoti kinų kalboje esančių žodžių dažnius, požymius. Toliau, dokumento lygiu, tyrimo rezultatai parodė, kad SVM metodo tikslumas buvo geresnis nei NB, o sukombinavus SVM su bigrama, buvo pasiektas geriausias klasifikavimo tikslumas [30]. Kitu atveju, kai buvo analizuojami telekomunikacijų klientų trumpieji komentarai (angl. *micro-blogs*), buvo nuspręsta nenaudoti dviejų populiariųjų metodų, o bandyti tikslumą pagerinti atsitiktinių miškų algoritmu (angl. *Random Forest*). Ansamblių segmento algoritmai tinkami kai turimuose duomenyse yra triukšmo arba kai norima išvengti modelio persimokymo. Šiuo metodu buvo sukurtas emocijas atskiriantis klasifikatorius, kuris sugebėjo suklasifikuoti sentimentus apie mobiliųjų telefonų prekės ženklus 83 proc. tikslumu [31].

Į neprižiūrimo mokymosi segmentą papuola visi leksikonu grįsti metodai ir sprendimai. Tai semantika grįsti metodai, kurie geba nustatyti sentimento poliariškumą iš kalbos taisyklių rinkinių, pvz.: žodynų ar neigiamą / teigiamą prasmę turinčių žodžių rinkinių, ir sukauptų kalbos euristikų. Šio metodo įgyvendinimui reikia sužymėti kiekvieną žodį ar frazę su atitinkamu poliariškumu, taip sukuriant žodynų rinkinius, išskiriant kalbines subtilybes, naudotinas frazes. Vienas pirmųjų tyrimų, kuriame panaudotas leksikono metodas, tai sukurtas sprendimas, kuris pirmiausia išskiria bigramas, atitinkančias tam tikras gramatines taisykles, įvertina jų poliariškumą, bei apskaičiuoja poliariškumo vidurkį visame dokumente [32]. Praktikoje pasitaiko, kad sukurti ir prieinami poliariškumą

nusakantys duomenų rinkiniai, netinkami analizuojamai tematikai, pvz., norima ištirti daug barbarizmų turinčią kalbą, tada tokiais atvejais generuojamas individualus žodynas pritaikytas tiriamąjai sričiai. Mokslininkai Araque, Zhu ir Iglesias [33], savo darbe pateikia 4 leksikono žodynus, rinkinius. Pirmasis, sukurtas Bing Liu, kuriame randami įvairūs teigiami, neigiami žodžiai, taip pat išskirti barbarizmai, neteisingi gramatiniai žodžiai, bei jie suklasifikuojami į teigiamo ir neigiamo poliariškumo segmentus. Kitas rinkinys, tai SentiWordNet, kuris yra plačiai naudojamas ir žinomas tyrėjų, rinkinys paremtas anglų kalbos baze ir medžio struktūra, o žodžiams nustatomas poliariškumas skalėje nuo 0 iki 1, pvz.: žodis lengvas, įgauna 0,625 teigiamus taškus ir 0,25 neigiamus taškus). Trečiąjį, emociniai standartai anglų kalboje (ANEW) rinkinį sudaro žodžiai, kuriems nustatytas emocinis balas, pagal žmonių reakcijas į tam tikrus žodžius. Bei ketvirtas, tai trečiojo modifikacija, kurioje papildomai integruojamas ir trumpuosiuose pranešimuose aptinkamas barbarizmas.

Prie leksikonu grįstų metodu, reikia pridėti ir kitus praktikoje taikomus algoritmus, kurie klasifikuojami kaip neprižiūrimo mokymosi. Tai visi algoritmai, kurių galutinė išvestis nėra konkrečiai apibrėžta, t. y. nežinoma kur tiksliai turi priklausyti stebėjimas. Šiai kategorijai priskiriami tokie algoritmai kaip klasterizavimo (k-vidurkių, k-medoidų ir kt.), dimensionalumo mažinimo (PCA ir kt.) ar kiti, tokie kaip anomalijų detekcija, ypatingųjų reikšmių dekompozicija (SVD). Šie algoritmai naudojami ir sentimentų analizėje, pvz.: dimensionalumo mažinimas gali būti naudingas, kai norima sumažinti vektorizuoto dokumento požymių skaičių, kai turima jų per daug. Klasterizavimo algoritmai gali būti naudingi norint automatiškai priskirti sentimento poliariškumą, t. y. priskiriant žymas dokumentams pagal jau anotuotą dokumentų dalį (angl. *semi-supervised*).

Praktikoje, norint pasiekti kuo geresnį klasifikavimo tikslumą, taikoma įvairių klasifikavimo metodų kombinacija. Kombinuoti modeliai dažniausiai įvardinami kaip hibridiniai. Pagrindinius hibridinių modelių principas, kad kiekviena modelio dedamoji, t. y. sub-modelis, geba geriau klasifikuoti atitinkamą dalį, pvz.: geba geriau klasifikuoti teigiamą poliariškumą kaupiančius dokumentus, kitaip sakant, sub-modelių išvestys - rezultatai negali koreliuoti. Galutinis sprendimas kokia dokumento sentimento orientacija, nusprendžiamas balsavimo principu. Panašus principas naudojamas ir modeliuose ansambliuose, pvz.: atsitiktinių miškų algoritmas sudarytas iš n sprendimų medžių, kur kiekvienas medis pateikia savo klasės klasifikavimo rezultatą, rezultatas suvidurkinamas bei gaunamas apibendrintas miško rezultatas dokumentui ar eilutei. Hibridiniai sprendimai apibūdinami kaip skirtingų algoritmų klasių pritaikymas.

Hibridinių modelių sentimentų analizėje taikymas - įprasta praktika. Ansamblio struktūra taikoma sentimentų klasifikavimo užduotims, siekiant efektyviai integruoti skirtingus požymių, kintamųjų rinkinius ir klasifikavimo algoritmus, kad būtų galima pagerinti prognozavimą. Tiriant filmų atsiliepimus, buvo pritaikytas Bejeso ir genetinių algoritmų hibridinis modelis. Kitas hibridinis sprendimas su „Twitter“ duomenis, buvo nagrinėtas frazės lygiu, kur taisyklėmis grįstais sprendimais buvo identifikuojami sentimentui darantys įtaką žodžiai, vėliau pritaikyta PCA išskirti įtakojančių žodžių požymiams bei klasifikavimui buvo naudojamas SVM algoritmas. Toks sprendimas leido nežymiai pagerinti tikslumą, nei naudojantis vien tik SVM [34].

Apibendrinant, krypčių atlikti sentimentų analizei yra ne viena. Nuo subjektyvios tyrėjo nuomonės ir ankstesnių tyrimų, galima pasirinkti kurį skirtinga klasifikavimo metodą naudotis ar remtis vien tik leksikonu grįstais metodais. Skirtingi metodai, duoda skirtingą tikslumą ir rezultatą, neretai, patys duomenys turi įtakos pasirenkant modelius. Nestruktūrizuotiems duomenims, pilniems barbarizmų, reikėtų rinktis metodus, kurie yra atsparesni triukšmui, o turint didelius masyvus duomenų, geriausia

naudoti modelius, kurie turi mažesnę persimokymo tikimybę. Įvairiais metodais gautus rezultatus, galima papildomai bandyti pagerinti pritaikant hibridinius sprendimus.

1.5. Sentimentų analize grįsti vartotojų patirties tyrimai

Nors sentimentų analizė pakankamai nauja tyrimų sritis, jos pritaikomas platus. Mokslininkų ar verslo atstovų dėmesys ir jų inicijuojami analizės procesai leidžia daryti prielaidas bei išsikelti pradines hipotezes. Įrodytą, kad internetiniais atsiliepimais išreikšta vartotojų patirtis turi įtakos įmonės pardavimus. Reikšminga teigiama koreliacija nustatyta tarp aukšto produktų įvertinimo ir pardavimų.

Mokslininkai Cali‘s ir Balaman‘as [18] kovo mėnesį publikavo mokslinį darbą, kuriame pateikė rekomendacinę sistemą, kuri grįsta internetinių atsiliepimų tyrimu, o veikimo principo pagrindas buvo sentimentų analizės būdu gauta informacija. Pirmiausia, leksikono metodu buvo išmatuojamas vartotojų pasitenkinimas, atsiliepimai nagrinėjami sakinio lygiu, nustatant sentimento poliariškumą. Kiekvienas atsiliepimas buvo priskiriamas tam tikrai abstrakčiai objektų grupei (pvz., personalas, aplinka ir kt.), toliau autoriai naudodami numatomų neapibrėžtųjų aibių teoriją (angl. *intuitionistic fuzzy set*), apskaičiuoja galimus alternatyvių prekių pasiūlymų balus. Šiais balais, panaudojus kitus kompleksiškus metodus, siūlomi produktai išrikiuojami. Ši mokslininkų sukurta metodika gali būti naudojama ir potencialiems klientams, kad jie galėtų įvertinti prekės ar paslaugos, kurią jie nori įsigyti, jau įsigijusių kitų klientų nuomones, bei papildomai gauti alternatyvius sprendimus. O iš įmonių pozicijos, nauda gaunama fiksuojant vartotojų pasitenkinimo lygį, kadangi bet kuriuo momentu, įmonė turi galimybę imtis papildomų veiksmų sumažinant nepasitenkinimą arba atvirkščiai – didinant pasitenkinimą. Tuo pačiu, verslas gali atsisakyti ir brangių, nuobodžių sprendimų, reikalaujančių daug žmogiškųjų resursų apdorojimui, tokių kaip Internetinė apklausos.

Kitame darbe analizuojama patirtis 2 skirtingais aspektais: vartotojų patirtis išreikšta per internetinius atsiliepimus, kurių poliariškumas buvo nustatytas sentimentų analize bei ekonominis rodiklis – bendras vartotojų pasitenkinimo indeksas. Šio darbo autorių tikslas buvo nustatyti ar yra koreliacija tarp vartotojų paliekamų atsiliepimų internetinėje erdvėje ir ar tai lemia bendrasis valstybės mastu išmatuotas vartotojų pasitenkinimas. Autorių siekis buvo surinkti kuo įmanomą didesnį masyvą įvairiakalbių duomenų, duomenys buvo gaunami „Twitter“, „Facebook“ socialinių platformų, forumų, blogų, „Youtube“ komentarų sekcijos. Analizuotoje Malaizijos rinkoje nustatyta koreliacija tarp šių kintamųjų labai maža bei kvestionuotino reikšmingumo. Autoriai daro prielaidą, kad viešosioms, valstybinėms ar kitoms suinteresuotoms įmonėms, kurios užsakinėja vartotojo pasitenkinimo indekso matavimus, pigiau, greičiau ir tikslingiau būtų galima vartotojo patirtį pamatuoti iš socialinių medijų gaunamos informacijos [19].

Rinkodaros specialistams svarbu įvertinti jų reklaminių kampanijų atsiperkamumą, pastebimumą, išmatuoti gautą naudą. Reklaminiai pranešimai vartotojui neretai būna erzinantys, per dažnai pasirodantys, kelia neigiamų emocijų ir kt. Būtent vartotojų emocijas ištyrė mokslininkė Tudoran‘a [20], kuri sentimentų analizės būdu išsiaiškino, kokiais motyvais remdamasis, reklamą pasiekiantis internetinių vartotojų segmentas, blokuoja reklaminius pranešimus. Atsiliepimų duomenys buvo surinkti iš įvairių naujienų, kurių kontekstas buvo reklamų blokavimas bei kitas su reklama susijęs turinys. Sentimentų analizės būdu buvo sugrupuotos nuomonės. Pastebėta, kad naujienos, kuriose dėmesys buvo skiriamas pačiai reklamai bei su reklama susijusiomis charakteristikomis, turėjo stiprų neigiamą poliariškumą, o naujienos, kuriose buvo kalbama arba siūloma paslauga įsigyti, prenumeruoti tinklapius ar kitas paslaugas be reklaminių pranešimų, sulaukė itin teigiamo sentimento

poliariškumo. Autorė tikina, kad toks nuomonių ištyrimas duoda nenuneigiamą naudą verslui, tinkamai interpretuojant gautus rezultatus, bei, netgi, gali identifikuoti verslui naudingos plėtros galimybes, pvz., už papildomą mokestį siūlyti bereklaminį tinklapį ar kt.

Sentimento analizėmis atlikti tyimai gali būti pritaikyta verslo praktikai, tiriant vartotojo nuomonę, taip išsiaiškinant problemines verslo sritis arba netgi identifikuojant naujas verslo plėtros galimybes.

1.6. Lietuviško teksto apdorojimo sprendimai ir nuomonės tyryba

Sentimentų analizės rezultatais grįsti sprendimai gali būti pritaikyti ir Lietuvos įmonėse sprendžiant įvairias problemas, tiriant vartotojus, analizuojant jų elgseną ir kt. Nors lietuvių kalboje turime gausų ir turtingą žodyną, tačiau dabartinės natūralios kalbos apdorojimo technikos ir sprendimai taip pat leidžia pakankamai tiksliai ištirti atsiliepimo sentimento poliariškumą lietuviškuose kanaluose, socialiuose tinkluose.

Tyrimų, kuriuose naudojami duomenys yra anglų kalba, gausu. Taip pat, nemažai skirtingų autorių tiria tuos pačius duomenų šaltinius, pvz., „Twitter“ trumpąsias žinutes. Suprantama, kad lietuvių kalba ir šios kalbos analizės nėra tokios populiarios ir maža dalis tyrėjų kuria sprendimus bei analizuoja lietuviškus tekstus. Pateikiami susisteminti ir apibendrinti darbai 1 lentelėje.

1 lentelė. Mokslinių tyrimų lietuviško teksto apdorojimo tematikoje apibendrinimas

Literatūros šaltinis	**Kryptis**	**Tikslas**	**Rezultatai**	**Tolimesnės perspektyvos**
Kapočiūtė-Dzikienė J., Krupavičius A. ir Krilavičius T. (2013) [35]	Sentimentų analizė / klasifikavimas	Išspręsti Lietuvos internetinių komentarų sentimento klasifikavimo užduotį, pritaikant 2 klasifikavimo metodus: žiniomis grįstą ir prižiūrimą mašininį mokymasi.	Pritaikius skirtingus duomenų rinkinio apdorojimo sprendimus (su, be išvalytais reikšmingais žodžiais, jaustukais ir kt.) bei pritaikius SVM ir Bajeso metodus, gauta, kad Bajeso metodas su unigrama ir bigrama kaip požymiais, klasifikavo geriausiai 0,697 tikslumu.	Atlikti išsamę klasių klaidų analizę, kuri padėtų rasti sprendimus kaip sumažinti klaidingą klasių klasifikavimą.
Petkevič V. (2018) [36]	Sentimentų analizė	Ištirti nuomonę, kuri buvo skleista socialiniuose tinkluose apie Lietuvos vyriausybę, prieš ir po Rusijos 2014 m. Krymo aneksijos.	Atlikus sentimentų analizę, pritaikius įvarius laiko eilučių metodus, nustatyta, kad po Krymo aneksijos, komentatoriai pozityviau pradėjo vertinti Lietuvos vyriausybę, jautėsi saugesni.	Patobulinti žodyną, pridedant gramatinių taisyklių bei pritaikyti mašininio mokymosi algoritmus, kurie gebėtų ištirti tekstų kontekstą.
Kapočiūtė-Dzikienė J., Damaševičius R. ir Wozniak M. (2018) [37]	Sentimentų analizė	Atlikti sentimentų analizę lietuviškiems internetiniams komentarams, naudojant gilaus mokymosi algoritmus: ilgąjį laikinosios atminties tinklą (LSTM) ir	Pritaikius skirtingas duomenų apdorojimo technikos variacijas, gauta, kad visais atvejais cNN metodas klasifikavo komentarus geriau nei LSTM, o pasiektas tikslumas sudarė 0.612.	Surinkti didesnį duomenų masyvą (šiuo atveju buvo tik 1500 dokumentų) ir vėl pabandyti gilaus mokymosi algoritmus.

Literatūros šaltinis	Kryptis	Tikslas	Rezultatai	Tolimesnės perspektyvos
		konvoliucinius neuroninius tinklus (cNN).		
Dorsch I., Nikolic J., Scheibe K., Zimmer F. ir Stock W. (2018) [38]	Sentimentų analizė	Nustatyti sentimento poliariškumą skirtingų šalių trumposiose žinutėse „Twitter“ tinkle, ištiriant taip pat ir Lietuvos naudotojus.	Tiriant Lietuvą visų šalių kontekste, nustatyta, kad Lietuva klasifikuojama kaip turinti žemiausią sentimento klimatą (0,071), t. y. sentimentai turi daug neigiamo poliariškumo.	Sentimento analizę atlikti su mašininio mokymosi algoritmais bei labiau ištirti patį sentimentą ir įtakojančius veiksnius.
Krilavičius T., Medelis Ž., Kapočiūtė-Dzikienė J. ir Žalandauskas T. (2013) [39]	Socialinių žiniatinklių analizė	Atlikti Lietuvos socialinių žiniatinklių stebėjimą, identifikuoti pagrindines diskusijų temas naudojantis informacijos gavimo (IR) ir natūralios kalbos apdorojimo (NLP) metodais.	Pritaikius teksto gavybos sprendimus ir gavus naujienų informaciją, buvo sėkmingai pritaikyti įvairūs natūralios kalbos metodai ant lietuviškų tekstų.	Vystyti darbus susijusius su NLP tematika.
Kapočiūtė-Dzikienė J. ir Damaševičius R. (2018) [40]	Kalbos apdorojimo sprendimas	Sukurti neuroniniais tinklais grįstą ir lietuviškos kalbos baze paremtą, žodžių įterpinių sprendimą.	Išbandytas skirtingas dimensionalumas nedavė reikšmingų rezultatų sprendimo įgyvendinimui, geriausias rezultatas pasiektas tęstinių žodžių krepšelio (CBOW) metodu. Geresnius rezultatus pasiekti trukdė kalbinės subtilybės, tokios kaip - sinonimai.	Pakartoti sprendimo įgyvendinimą su didesne duomenų imtimi.

Verta paminėti, kad lietuviškų tekstų sentimento tyrimai bei analizės kelia susidomėjimą ir tarp aukštojo mokslo studentų bei busimųjų mokslininkų. Analizuojant literatūrą, susijusią su sentimento analize, rasti 6 viešai prieinami magistro baigiamieji projektai, kurie parašyti per pastaruosius keletą metų bei kurių tematika atitinka šio projekto temą.

Patvirtinama prielaida, kad lietuviškų tekstų analizė yra dar nauja sritis, tačiau vis populiarėjanti tyrėjų tarpe, kadangi didelė dalis literatūros yra iš praėjusių metų. Ši tema aktuali, kadangi nerandama darbų, kur būtų plačiau palyginama įvairi metodologija, išbandomi hibridiniai sprendimai bei palyginamos skirtingos teksto vektorizavimo technikos. Iš įmonių pozicijos, darbų, kur sentimentų analizė buvo pritaikyta gerinant vartotojų patirtį ar bent įmonės procesus susijusius su vartotojais, nebuvo rasta visai. Keleta rastų darbų sentimentų analizės tematika, analizavo interneto vartotojų komentarus išreikštus įvairiomis tematikomis, pvz., atsakas į naujieną ar tiesiog nestruktūrizuotas nuomonės reiškimas nelabai susietinam kontekstui. Šis magistro projektas bus indėlis į lietuviškų tekstų analizę bei praktišką sentimentų analizės pritaikymą, tiriant lietuviško verslo klientų atsiliepimus, paliktus internetinėje erdvėje. Projekte naudojamas duomenų šaltinis - „Facebook“ socialinio tinklo įmonių rekomendacinių atsiliepimų skiltyje palikti vartotojų atsiliepimai, išvis

nebuvo giliau tirti pasauliniu mastu, kadangi „Facebook“ API tokio pasirinkimo nesiūlo, dėl duomenų saugumo užtikrinimo.

2. Tyrimų metodai

Šioje darbo dalyje apžvelgiami visi tolimesniame tyrime naudojami natūralios kalbos apdorojimo metodai, mašininio mokymosi algoritmai ir tikslumo įvertinimo metrikos. Pirmiausiai, aprašomi duomenų vektorizavimo metodai: DBoW, DBoWs, DMPV, LSI, LDA, RP, Sent2Vec (fastText), BERT. Toliau, mašininio mokymosi algoritmai: logistinė regresija (LogR), SVM, RF ir XGBoost (Extrieme Gradient Boosting). Tikslumas bus vertinamas ROC, DET, PR kreivėmis, universaliu tikslumo matu AUC bei matais iš sumaišymo matricos F-matu ir Kappa.

2.1. Duomenų vektorizavimo metodai

Dauguma mašininio mokymosi algoritmų reikalauja, kad duomenų įvestis būtų fiksuoto ilgio požymių vektorius. Jeigu nagrinėjami tekstai, vienas iš populiariausių metodų – tai žodžių krepšelis (angl. *bag of words*), bei šio metodo modifikacijos, kiti semantika grįsti modeliai. Taip pat, keletas metodų panaudoti iš pastraipų vektoriaus segmento, kur modeliai gali pakankamai tiksliai įvertinti semantinį ryšį tarp skirtingo ilgio tekstinių dokumentų, pvz.: žodžių, sakinių, pastraipų.

Duomenų vektorizavimas atliekamas turint dokumentą, sakinį ar teiginį, šis turimas objektas apdorojamas pasirinktu metodu bei gaunama nustatyto dimensionalumo išvestis. Dimensionalumas yra subjektyvus dalykas, todėl priklauso nuo tyrėjo pasirinkimo, taip pat praktikoje neretai įvertinamas skirtingų dimensionalumų tikslumas. Išvestis gaunama skaitinėje formoje. Mašininio mokymosi algoritmai tokią išvestį skaito kaip požymių rinkinį.

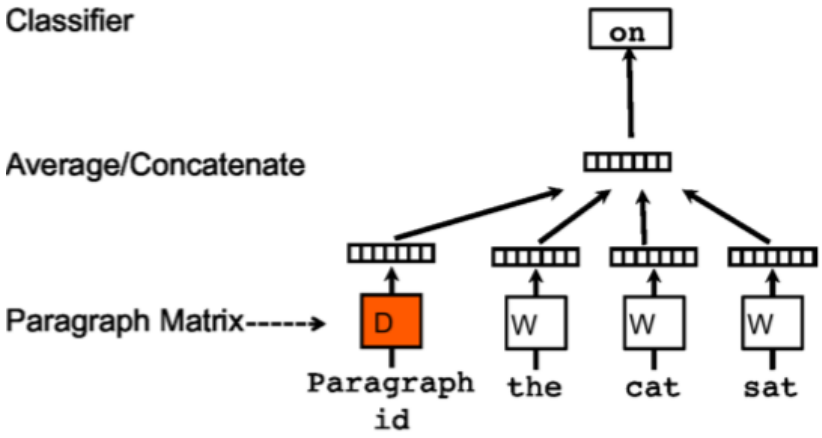
2.1.1. Paskirstytos atminties modelis

Tai pastraipų vektoriaus modelis, kur pastraipos požymis gali būti traktuojamas kaip žodis. Šis metodas veikia kaip atmintis, t. y. metodas geba prisiminti ko trūksta duomenų įvesties kontekste arba geba atsiminti pastraipos tematiką, dėl šios priežasties šis metodas vadinamas atminties modeliu. Šis modelis yra apmokytas naudojantis stochastiniu gradientinio nusileidimo metodu, kai gradientas gaunamas naudojant atgalinį sklidimą. Kiekviename stochastinio nusileidimo etape, galimas tik fiksuoto ilgio kontekstas iš atsitiktinės pastraipos.

Tarkime, bazėje yra N pastraipų ir M žodžių žodyne. Tikslas yra apmokyti pastraipų vektorius taip, kad kiekviena pastraipa turėtų p dimensijų, o kiekvienas žodis q dimensijų. Tokio modelio parametrų skaičius randamas pasinaudojus formule:

$$N \times p + M \times q; \quad (1)$$

Atvejais, kai turimi dideli masyvai informacijos, turimų parametrų skaičius išauga ir pasidaro kompleksiškas.

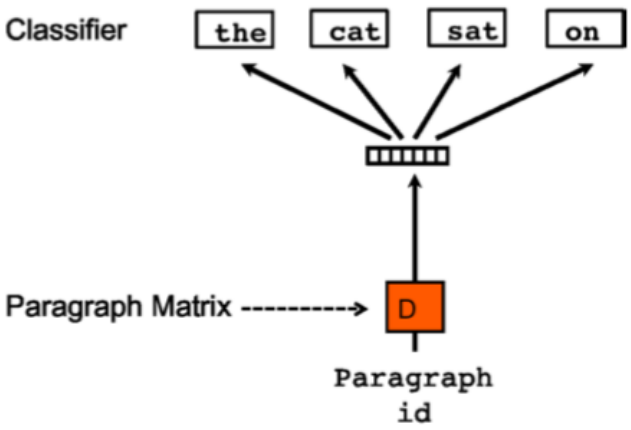


5 pav. Paskirstytos atminties modelis

Paveiksle 5, atvaizduojama pastraipų vektoriaus pagrindu sukurtas paskirstytos atminties modelio veikimas. Pasinaudojus vektoriaus vidurkiu, kurio kontekstas sudarytas iš 3 žodžių, bandoma prognozuoti ketvirtąjį. Pastraipų vektorius atvaizduoja trūkstamą informaciją, o atminties įgalinimas padeda prisiminti pastraipos tematiką. Matoma, kad kiekviena pastraipa atvaizduojama unikaliu vektoriumi matricoje D, o žodžiai vektoriuje W. Toliau, pastraipų vektoriai suvidurkinami, bei gaunamas pastraipą reprezentuojantis požymių vektorius, kuris gali būti naudojamas mašininio mokymo algoritmų įvestyje [41].

2.1.2. Paskirstyto žodžių krepšelio modeliai

Paskirstytos atminties modelio paprastesnė ir mažiau kompleksiška modifikacija, tai paskirstytas žodžių krepšelio metodas (DBoW). Šis metodas taip pat priskiriamas prie pastraipų vektoriaus metodų. Šio metodo principas, kad įvesties kontekste esantys žodžiai neįvertinami, tačiau bandoma atsitiktiniu būdu prognozuoti tai iš pastraipos. Kiekvieną kartą, kai inicijuojamas stochastinio gradientinio nusileidimo metodas, atsitiktiniu būdų sugeneruojama tekstinė erdvė, tada pasirenkamas žodis ir pagal pastraipos vektoriaus taisykles prognozuojama reikšmė [41]. Šio modelio veikimo principas pateiktas 6 paveiksle.



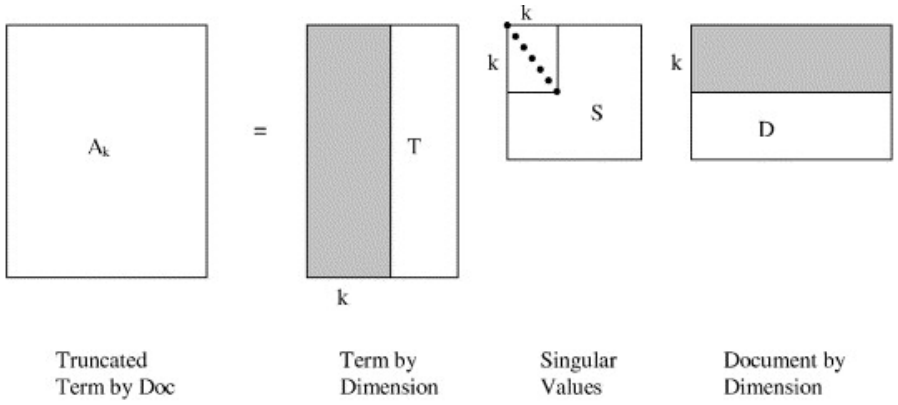
6 pav. Paskirstyto žodžių krepšelio modelis

Panašiu algoritmo principu sukurtas ir „Skip-gram“ metodas. (DBoWs). Šis metodas bando prognozuoti žodžius aplink vektoriaus centrą, išskirdamas tuos, kurie lemia pastraipos (dokumento)

kontekstą. Šiame metode, kiekvienas žodis reprezentuoja žodžių vektorių W, o kiekvienas kontekstas reprezentuoja konteksto vektorių C, kur d pateikiamas kaip vektoriaus dimensionalumas. Atliekamas vidinis dauginimas tarp W ir C vektorių ir apmokyto tinklo išvestyje gaunami sluoksniai su žodžių svoriais. Šis metodas geba tiksliai išskirti didelio dimensionalumo vektorius netgi iš didelių masyvų duomenų, bei išskirti svarbius žodžius, kurie nėra populiarūs [42].

2.1.3. Latentinis semantinis indeksavimas

Latentinis semantinis indeksavimas, tai populiarus informacijos gavybos algoritmas. LSI gali būti pritaikytas sprendžiant įvairias užduotis: paieškos ir gavybos, klasifikavimo, filtravimo. Šis metodas tai vektorinės erdvės sukūrimo metodas. Metodo pagrindas yra singuliarių reikšmių dekompozicijos (SVD) algoritmas. SVD algoritmas padalina matricą į T – terminų, S – singuliarių reikšmių ir D – dokumentų matricų projekcijas, kur kiekviena reprezentuoja matricos elementų struktūrą.



7 pav. SVD metodu gautos projekcijos

Esant LSI sistemoje, T, S ir D matricos yra sumažinamos iki k dimensionalumo. Tamsiai pilkos zonos matricose T ir D yra paliekamos, o ne pilkos – pašalinamos. Toks išskaidymas padeda atsisakyti triukšmo latentinėje ervėje ir gaunamos reikšmingų žodžių sąsajos, kur panašumas apskaičiuojamas kosinuso panašumo matu [43].

LSI metodas naudoja terminų – dokumentų matricą, kurios esmė yra perteikti sutinkamų terminų dažnį dokumentuose. Tokia matrica gali būti konstruojama pasinaudojus termino dažnio – atvirkštinio dokumento dažnio (TF-IDF) metodą. Svorių suteikimas šiuo metodu tai dažnai naudojama procedūra teksto tyryboje ir informacijos gavyboje, su tikslu nustatyti lingvistinių terminų svarbą analizuojamoje kalbinėje bazėje. Taikant šį metodą, susiduriama su žodžio krepšelio algoritmu, kurio pagalba yra suskaičiuojamas žodžių dažnis dokumentuose. Termino svarbumas (svoris) padidėja kiekvieną kartą terminui atsirandant nagrinėjamuose tekstuose, tačiau dažnai surandamų ir nereikšmingų žodžių, pvz., ‚už‘, ‚tada‘, svoriai nuo jų per didelio dažnumo viename dokumente, yra sumažinami. Turint terminus t, kurie sudaro terminų visumą, kuri sutinkama erdvėje N dokumentų d, svoris randamas pritaikius formules:

$$\begin{aligned} \mathrm{tf}_{t,d} &= \frac{f_{t,d}}{n_d}; \\ \mathrm{idf}_t &= \log\frac{N}{df_t}; \\ W_{t,d} &= \mathrm{tf}_{td} \times \mathrm{idf}_t; \end{aligned} \tag{2}$$

Čia - $f_{t,d}$ terminų dažnis dokumente, df_t yra termino dokumentų dažnis.

Sprendžiant klasifikavimo problemą, pvz.: taikant sentimentų analizę, dokumentas pakeičiamas į sentimento klasę (teigiamas / neigiamas), tada termino dažnis yra skaičiuojamas klasės lygiu, o ne dokumento, o atvirkštinis dokumentų dažnis tampa, atvirkštiniu komentarų, atsiliepimų dažniu. Kur N tampa komentarų erdve, o terminų dažnis dokumente skaičiuojasi šioje erdvėje [44]. Dėl gaunamų išretintų, didelio dimensionalumo duomenų, papildomai integruojamas PCA metodas dimensijų sumažinimui.

2.1.4. Latentinis Dirichlė paskirstymas

Šis metodas, taip pat kaip ir LSI, pagrįstas stochastiniu gradiento optimizavimu. LDA tai tikimybinis Bajeso modelis tekstiniams dokumentams. Daugiausiai šis metodas naudojamas norint iš dokumentų išgauti juose dominuojančias tematikas. Tarkime, turimas k tematikų skaičius, kur $k \in \{1, \ldots, K\}$. Visas tematikų rinkinys sudaro žodyną, kuris apibrėžtas tikimybių vektoriumi $\beta_k \sim \text{Dir}(\eta)$. Atsižvelgiant į temas, LDA algoritmas kiekvienam dokumentui įvykdo generacinį procesą. Pirmiausia, nubraižomas tematikų pasiskirstymo tikimybių vektorius $\theta_d \sim \text{Dir}(\alpha)$ ir kiekviename dokumente esančiam žodžiui sugeneruojamas tematikos numeris $zdi \in \{1, \ldots, K\}$, pagal tematikų svorius bei pateikiami tematiką reprezentuojantys žodžiai.

Norint pritaikyti LDA dokumentų rinkiniui, pirmiausia išnagrinėjamas aposteriorinis skirstinys pagal tematikas β, tematikų proporcijas θ ir tematikų priskyrimo z taisykles. Taip identifikuojama latentinė duomenų rinkinio struktūra, kuri toliau gali būti naudojama prognozuojant ar apžvelgiant duomenis. Aposteriorinis skirstinys negali būti išmatuojamas tiesiogiai, todėl tikėtinam skirstiniui sukurti naudojami Monte Carlo simuliacijos metodai [45].

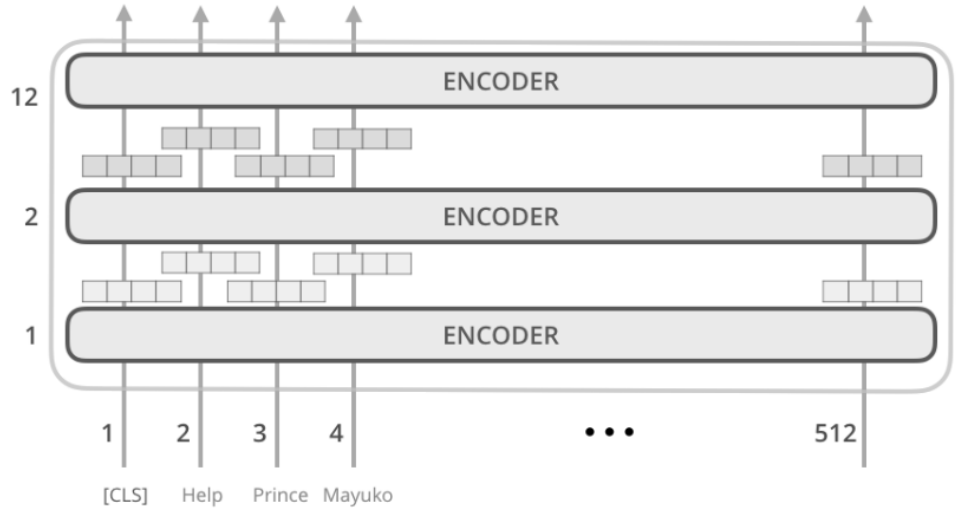
2.1.5. Atsitiktinis indeksavimas

Vektorinė semantinė analizė, tai technologija, leidžianti gauti semantiškai panašius terminus, frazes iš tekstinių duomenų. Dar vienas tekstinių duomenų vektorizavimą įgalinantis metodas – atsitiktinės projekcijos arba atsitiktinis indeksavimas. Šis metodas naudoja išretintus (angl. *sparse*), didelio dimensionalumo atsitiktinius indeksuotus vektorius dokumento atvaizdavimui. Kiekvienas dokumentas yra priskiriamas prie atsitiktinio indeksuoto vektoriaus, o terminų, frazių panašumas apskaičiuojamas pasinaudojant kontekstinių ryšių matricą (angl. *terms-by-contexts co-occurrence matrix*). Šioje matricoje terminai yra eilučių vektoriai, kurių dimensionalumas išlieka toks pat kaip ir atsitiktinių vektorių. Dokumente aptikus terminą, dokumento atsitiktinio indeksavimo vektorius pridedamas į atitinkamą eilutę. Taip visi terminai sudaro matricą, kurioje matomas kiekvieno termino pėdsakas bendroje semantinėje struktūroje [46].

2.1.6. BERT

Vienas naujausių teksto vektorizavimo metodų – BERT (angl. *Bidirectional Encoder Representations from Transformers*), buvo pristatytas 2018 m. „Google“ kompanijos dirbtinio intelekto tyrėjų grupės. BERT pritaikomumas yra platus, turint vektorizavimo užduotį ir norint gauti tekstinio sakinio fiksuoto skaičiaus dimensionalumą, pirmajam tokenui, žodžiui ar kitam duomenų imties reprezentatyviam objektui (angl. *token*), kuris pagal konstrukciją atitinka specialų žodžių įterpinį [CLS], imama paskutinė būsena, pvz., išvestis. Pirmasis sakinio tokenas visada būna specialus

klasifikavimo įterpinys ([CLS]). Paskutinis sluoksnis visada turi sąryšį su šiuo tokenu, kuris naudojamas kaip sakinio reprezentatorius klasifikavimo užduotyse [47].

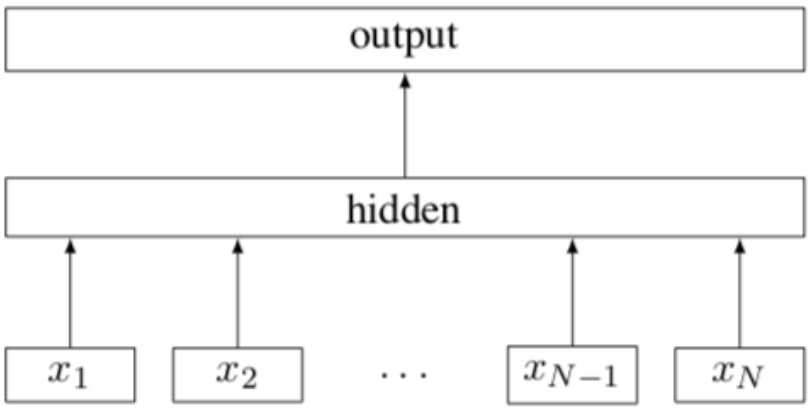


8 pav. BERT loginė schema

BERT metodu gaunama tiek sluoksnių, koks BERT metodas yra naudojama. Kiekviename kodavimo (angl. *encoder*) sluoksnyje, turimi tokeną reprezentuojantys duomenys, gali būti panaudojami kaip tokeno požymiai, dažniausiai pasirenkamas paskutinis sluoksnis. Kuriant mašininio mokymosi modelius, BERT metodu gauta duomenų išvestis turės būti dar papildomai modifikuota, kiekvieną tokeną reprezentuojantis vektorius, turės būti sumažintas bei agreguojamas iki dokumentą reprezentuojančio vektoriaus.

2.1.7. FastText (Sen2Vec)

Tai metodas, kuris buvo pristatytas „Facebook" dirbtinio intelekto tyrėjų grupės. Šio metodo pranašumas nuo kitų metodų, pvz., pastraipų vektoriaus metodų, tai kad metodas geba priimti ne pilną žodį, o žodį išskaido į n-gramu elementus, kur n svyruoja nuo 1 iki žodžio elementų skaičiaus. Šiuo metodu galima aptikti itin retus ir didelę informacinę vertę turinčius žodžius, tuo pačiu, modelis sugeba ir toliau veikti ir vektorizuoti žodžius kurių nėra sudarytame žodyne bei yra atsparus įvairioms žodžio kalbinėms formoms.



9 pav. FastText loginė schema

Šis metodas yra sukurtas neuroninių tinklų pagrindu, kuriame yra vienas sluoksnis. Kur pirmiausia, į paieškos sluoksnį įkeliami dokumentai žodžio n-gramų lygiu, kur kiekviena n-grama vektorizuojama iki žodžio. Toliau, visi žodžiui gauti n-gramų vektoriai yra suvidurkinami, kad būtų gaunamas vienas vektorius identifikuojantis žodį o vėliau - dokumentą. Paslėptame sluoksnyje yra įterpiami pasirinktieji parametrai, tokie kaip vektoriaus dimensionalumas ar žodyno dydis.

2.2. Klasifikavimo modeliai

Statistikoje ir mašininio mokymosi vystymo srityse, teksto klasifikavimo uždavinys, tai tekstinių dokumentų priskyrimas iš anksto nustatytam kategorijų rinkiniui. Šis procesas yra tarpdisciplininis, kadangi apima skirtingas sritis kompiuterių moksle: NLP, ML ir statistikos teoriją. Konkrečiai šiame projekte klasifikavimo užduotis specifikuojama iki detekcijos uždavinio, kadangi modeliavimo rezultato klasės bus binarinės, dėl šios priežasties klasifikavimo algoritmai bus paprastesni ir, tikimasi, generuos tikslesnius rezultatus.

Mašininio mokymosi algoritmai yra taikomi daugelyje sričių. Praktika rodo, kad SVM algoritmas yra linkęs geriau prognozuoti kai egzistuoja didelis duomenų dimensionalumas turint didelį duomenų rinkinį, o RF ir XgBoost turi savo pliusų dėl persimokymo problemos eliminavimo. Keleto skirtingų algoritmų pritaikymas leis įvertinti kurio tikslumas geresnis turint projekte naudojamą duomenų rinkinį.

2.2.1. Logistinė regresija

Logistinės regresijos veikimas panašus į klasikinę, daugelyje sričių taikomą tiesinę regresiją, tačiau logistinėje regresijoje įvesties priklausomas kintamasis yra binominis. Rezultatas gaunamas iš visų modelio lygties nepriklausomų kintamųjų įtakos, įvertinant jų ryšio stiprumą. Logistinė regresija yra vienas populiariausių ir mažiausiai kompleksiškų algoritmų, todėl šio algoritmo vystymas, modifikacijos yra dažnos tyrėjų tarpe. Paprastas logistinės regresijos modelis aprašomas

$$\text{logit}(Y) = \ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta X; \tag{3}$$

$$\pi = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}; \tag{4}$$

čia π tikimybė, kad atsitiks įvykis; α – laisvasis narys; β – regresijos koeficientai; e – matematinė konstanta, kuri yra pagrindas natūraliai logaritmo funkcijai, $e \approx 2.71828$.

X nepriklausomi kintamieji gali būti kokybiniai arba tolydieji, ryšys tarp logit(Y) ir X yra tiesinis. Tačiau X ir Y tikimybės ryšys netiesinis, dėl šios priežasties reikalinga logaritmo transformacija. Regresijos koeficientai apibūdina ryšio kryptį tarp X ir logit(Y). Kai koeficientas teigiamas, galima daryti prielaidą, kad nepriklausomas kintamasis sąveikauja teigiamų ryšių, kai koeficientas yra didesnis už 0. Hipotezių tikrinime, jeigu regresijos koeficientas yra 0, daroma prielaida, kad ryšio tarp atitinkamo nepriklausomo kintamojo ir priklausomo kintamojo logaritmo – nėra. Jeigu nepriklausomas kintamasis, kuriam nustatytas ryšys, yra binarinis, tada jo ryšio stiprumą nusakantis dydis (angl. *odds ratio*) yra matematinė konstanta e, bei gali būti išreikšta kaip $\beta(e\beta)$. Kai į modelį norima įtraukti daugiau nepriklausomų kintamųjų, modelio lygtis pasidaro kompleksiškesnė ir gali būti išreikšta universalesniu apibrėžimu [48].

$$\text{logit}(Y) = \ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n; \quad (5)$$

čia n – nepriklausomųjų kintamųjų skaičius.

2.2.2. Atsitiktiniai miškai

Mokslininkas Breiman‘as savo tyrimuose (1996, 2000, 2001, 2004), parodė, kad klasifikavimo ir regresijos didesnis tikslumas gali būti pasiektas pasinaudojus ansamblių modelius. Jis įgalino sprendimų medžių metodą, kuris buvo pavadintas atsitiktiniais miškais. Galutinis rezultato sprendimas paremtas visų modelyje esančių medžių balsavimo vidurkiu. Atsitiktinių miškų metodas yra greitas ir lengvai pritaikomas, šio metodo rezultatų išvestis pasižymi aukštu tikslumu. Turint didelės apimties duomenų įvestį, dėl kiekvienos modelio komponentės (medžio) įtakos sprendimui, išvengiama persimokymo problemos [49].

Atsitiktinis miškas pirmiausia pradedamas formuoti pagal pasirinktą medžių skaičių, kur medžių požymiai priskiriami atsitiktiniu būdų, toliau kiekvienos iteracijos metu, požymių reikšmingumas perskaičiuojamas, paliekant geriausiai duomenis reprezentuojančius medžių požymius. Medžiai auginami pasinaudojus CART metodologija. Medžių auginimas turėtų būti baigtas, pasiekus ribinę dalį, kur modelio rezultatų tikslumui papildomų medžių kiekis neteikia reikšmingos įtakos prognozavimo tikslumo matui (OOB). Nustatant OOB mato dydį, įvertinami tik tiek medžiai, kurių sudarymui nebuvo naudota apmokymo imtis, t. y. atsitiktinių miškų metodas pasilieką dalį duomenų iš visos duomenų įvesties, tikslumo vertinimui.

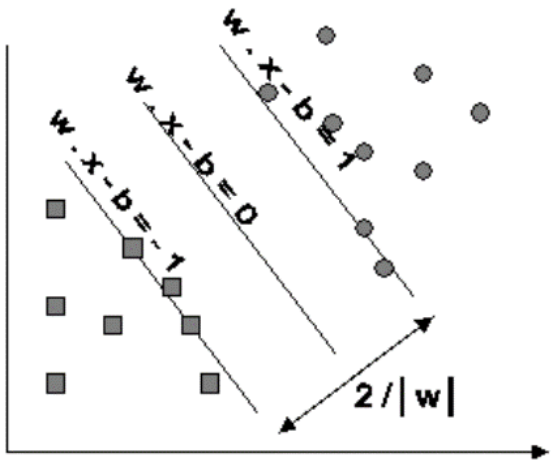
Norint prognozuoti x stebinio klasę, sukuriamas naujas RF, kurio sudarymui naudojami tik tie medžiai, kurie teisingai klasifikuoja testavimui skirtus duomenis. Visų prognozuojamų klasių suminis svoris lygus 1. Modelio apmokymo metu, klasių svoriai yra proporcingai išskirstyti lygiomis dalimis. Tuo atveju, kai yra klasių disbalansas, klasių svorinis koeficientas apskaičiuojamas naudojantis formule [50]:

$$w^{(k+1)}(n) = \frac{w^k(n)\beta_k^{d(n)}}{\Sigma w^{(k)}(n)\beta_k^{d(n)}}; \quad (6)$$

Čia $w^{(k)}(n)$ yra k žingsnio svoris, β_k – klaidingas svorių vertinimas.

2.2.3. Atraminių vektorių metodas

Vienas populiariausių klasifikavimo metodų, kuris pamėgtas tyrėjų dėl itin tikslaus rezultato. Tai prižiūrimo mašininio mokymosi metodas, kuris naudojamas klasifikavimui ir regresijai. Metodo veikimas grįstas paklaidos minimizavimu ir paraščių, kurios atskiria klases modeliuojamoje erdvėje, maksimizavimu. Kitaip SVM metodas dar vadinamas paraščių maksimizavimo metodu. Metodas modeliuoja įvesties vektorių į multidimensionalumo erdvė, kurioje įgalinta klases skirianti hiperplokštuma. Esant didesniam atstumui tarp hiperplokštumos paraščių, klasifikatorius geba tiksliau atskirti klases.



10 pav. SVM tiesinis modelis

Paveiklas atvaizduoja sukurtą hiperplokštumą su maksimaliomis 2 klases atskiriančiomis paraštėmis. Taškai esantys ant paraščių, vadinami atraminiais vektoriais. Apmokymo duomenų vektorius pagal funkciją Φ yra integruojamas į nustatyto dimensionalumo erdvę. Pagal subjektyviai pasirinktą branduolio tipą, apskaičiuojama branduolio funkcija, projekte bus naudojama tiesinė modifikacija [51]. Tiesinės modifikacijos lygtis:

$$K(x_i, x_j) = x_i^T x_j; \tag{7}$$

čia x_n yra p-dimensijos realus vektorius.

Tiesinio SVM modelio vystymui pagrindinis dėmesys skiriamas į baudos parametrą C. Tiesinis SVM metodas gali būti su minkštomis arba kietomis paraštėmis, praktikoje visada taikoma minkštų paraščių modifikacija, tačiau papildomai įvedamas baudos parametras C, kuris taikomas duomenims kirtus paraščių erdvę.

2.2.4. Gradientinis stiprinimas

Gradientinis stiprinimas (toliau XGBoost), tai mašininio mokymosi metodas skirstas spręsti klasifikavimo ir regresijos užduotis. Panašiai kaip ir atsitiktinių medžių metodas, visas modelis susideda iš kitų mažesnių modelių - klasifikatorių, kurie geba geriau prognozuoti atitinkamus duomenų požymius. Komponentiniai modeliai į bendrą klasifikatorių įtraukiami ir tada, kai reikia naujai prognozuoti ankstesnių modelių paklaidas. Metodas pavadintas gradientiniu stiprinimu (angl. *gradient boosting*) kadangi naudojamas gradientinio nuolydžio metodas optimizuoti nuotolio funkciją, kai įvedamas naujas modelis. Modelio išraiška:

$$\hat{y}_i = \sum_{k=i}^{K} f_k(x_i), f_k \in K; \tag{8}$$

čia K yra medžių skaičius, f yra funkcinės erdvės F funkcija, F visų įmanomų medžių rinkinys.

Modelio tikslas minimizuoti tikslo funkciją, kuri gali būti aprašoma:

$$obj(\theta) = \sum_i^n l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k); \tag{9}$$

čia $l(y_i, \hat{y}_i)$ yra apmokymo imties nuotolio funkcija, $\Omega(f_k)$ reguliarizavimo taisyklės.

Kadangi nėra lengva apmokyti visus medžius iškarto, jie apmokami pažingsniui, žingsnyje t prognozuojama reikšmė išreiškiama formule [52]:

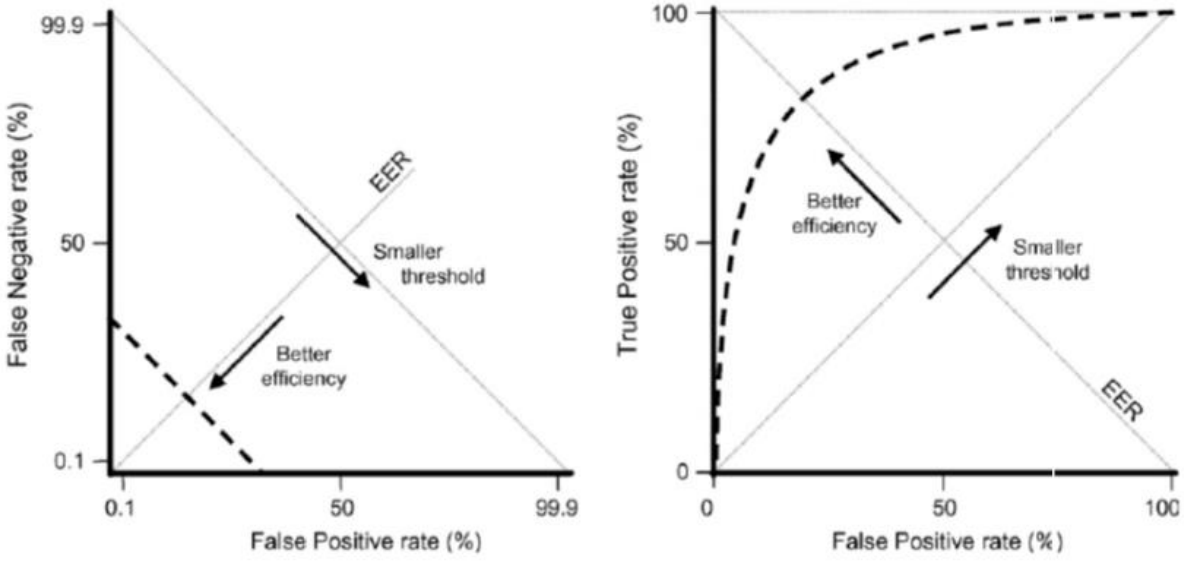
$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i); \quad (10)$$

Modelis yra pamėgtas srities tyrėjų dėl didelio tikslumo, išvengiama persimokymo problemos, modelis yra lankstus dėl galimų subjektyvių parametrų pakeitimo ir nustatomų reguliarizacijos parametrų.

2.3. Tikslumo vertinimas

Norint įvertinti sukurto modelio kokybę ir rezultatų tikslumą, daugumoje užduočių, kuriose naudojami mašininio mokymosi algoritmai, skaičiuojamos tikslumo metrikos, braižomi tikslumą įvertinantys grafikai. Tikslumas vertinamas ir norint tarpusavyje palyginti keletos naudojamų algoritmų sukurtų modelių tikslumą, išsirenkant geriausiai atliekantį nustatytą užduotį. Egzistuoja ir universalūs matai, kurie tinkami daugumoje užduočių, bei specifiniai naudojami tik tam atitinkamose užduotyse. Šiame projekte turimas klasifikavimo uždavinys, kadangi klasifikuojamos tik 2 klasės, uždavinys specifikuojasi į detekcijos uždavinį. Detekcijos uždavinio tikslumui išmatuoti yra naudojama viena plačiausių tikslumo matavimo matų aibė.

Tikslumui nustatyti gali būti naudojamos vaizdinės tikslumą reprezentuojančios priemonės – grafikai. Viena iš informatyviausių kreivių – ROC. Šios kreivės paskirtis, atvaizduoti sukurto klasifikatoriaus jautrumo ir specifiškumo ryšį. Šią kreivę galima pritaikyti palyginant skirtingus klasifikatorius, tada klasifikatorius, kurio kreivė yra aukščiausiai atsitiktinio spėjimo tiesės (tiesė einanti įstrižai), interpretuojamas kaip geriausias. Plotas, esanti po ROC kreive, tai AUC matas. Šis matas padeda identifikuoti kaip tiksliai klasifikatorius sugeba atpažinti tiriamo objekto klasę, jeigu AUC < 0.5, galima teigti, kad sukurtas klasifikatorius prognozuoja blogiau nei atsitiktiniai spėjimai.

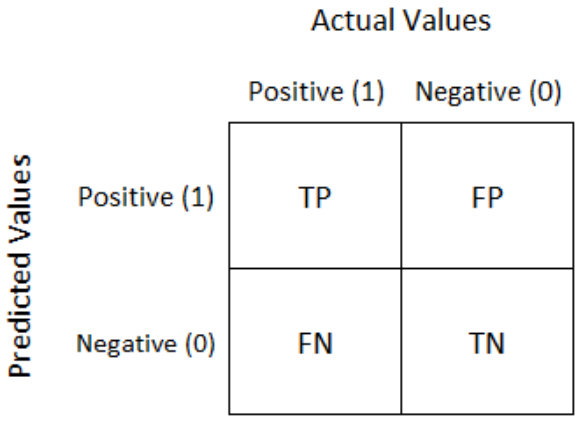


11 pav. DET (kairėje) ir ROC (dešinėje) kreivės [54]

Kita kreivė - DET, panaši į ROC, tačiau ordinačių ašyje pateikiamas ne teisingai suklasifikuotos teigiamos klasės, o neteisingai klasifikuotos neigiamos klasės reikšmės. DET kreivės atveju, geresnis modelis yra tas, kurios kreivė yra žemesnė. Taip pat, informatyvus gali būti ir taškas, kuriame klasifikatoriaus klaidų kiekis yra lygus – EER (angl. *equal error rate*), kuris padeda nustatyti ribinį

klasifikavimo tašką ir papildomai peržiūrėti kodėl atitinkami duomenys yra neteisingai klasifikuojami. Kuo šis taškas mažesnis, tuo prognozavimo tikslumas didesnis.

Prie EER slenksčio gaunami įverčiai, reikalingi nubraižyti informatyvius grafikus. Šiems įverčiams gauti ir klasifikavimo modelio veikimui įvertinti, naudojama sumaišymo matrica. Sumaišymo matrica, tai lentelė, kurioje pateikiamas sukurto modelio sąsajos ir išmatuoti įverčiai, gauti ant nematytų, testavimui skirtų duomenų. Kitaip sakant, norint pamatyti kaip klasifikavimo modelis veikia ant dar nematytų duomenų, sumaišymo matricos pateikimas yra vienas informatyviausių būdų kadangi matomas teisingai ir neteisingai klasifikuotų klasių santykis.



12 pav. Sumaišymo matricos logika

Sumaišymo matricoje matyti kokios buvo tikrosios duomenų klasės ir kokias modelis prognozavimo. Galima pastebėti ar modelis geriau prognozuoja vieną ar kitą klasę ir nuspėti, pvz., egzistuojančią klasių disbalanso problemą. Turint prognozavimo tikslumo duomenis, galima apskaičiuoti duomenų prognozavimą reprezentuojančius matus: tikslumą (angl. *precision*), atkuriamumą (angl. *recall*) bei šių dviejų matų kombinaciją – F-matą (angl. *F score*). Šios metrikos įvertina tokius kriterijus kaip teisingi rezultatai (TP, TN), netikėti rezultatai (FP), trūkstami rezultatai (FN). Tikslumas apskaičiuojamas

$$\frac{TP}{(TP+FP)}; \tag{11}$$

o atkuriamumas

$$\frac{TP}{(TP+FN)}; \tag{12}$$

Atkuriamumo ir tikslumo kintamumas lyginant skirtingus modelius, gali būti pateiktas ir grafiškai. Atvaizduoti šiuos 2 matus galima nubraižant juos koordinačių sistemoje, kur tikslumas atidedamas ordinačių ašyje, o atkuriamumas abscisių ašyje. Kreivė parodo, kiek modelis gerai prognozuoja teigiamą klasę.

Projekte naudojamas F-matas, tai pusiausvyra tarp šių dviejų matų. Šis matas padeda įvertinti modelių tikslumą geriau nei pastarieji matai, ypač tais atvejais, kai yra nevienodas klasių pasiskirstymas (klasių disbalansas). Kitas įvertinimo matas gaunamas iš sumaišymo matricos - Kappa. Šiuo matu projekte bus vertinamos vektorizavimo ir modelių kombinacijos tarpusavyje, išrenkant geriausiai klasifikuojantį, kai duomenys būna ne visada tokie patys ir paprasti tikslumą vertinantys matai tampa

nereikšmingi. Kappa vertina numatomą tikslumą su tikėtinu tikslumu, kur numatomas tikslumas tai visi teisingai suklasifikuoti atvejai iš sumaišymo matricos, o tikėtinas tikslumas apibrėžiamas kaip kiekvieno atsitiktinio klasifikatoriaus galimas pasiekti tikslumas. Šis matas kinta rėžiuose tarp 0 ir 1, kur 1 reiškia pasiektą idealiausią tikslumą.

2.4. Triukšmo šalinimas

Turint klasifikavimo uždavinius, požymių ir klasių triukšmas yra kokybę įtakojantis veiksnys. Triukšmas esantis kintamuosiuose įtakoja duomenų požymius, o triukšmas esantis klasėse gali pakeisti duomenų eilutėje priskirtą klasę. Tarkim, teigiamą klasę turinti duomenų eilutė, dėl triukšmo gali būti klasifikuojama kaip neigiama. Triukšmas gali būti traktuojamas kaip anomalija, išskirtis, Gauso kreivės kraštutinumas. Anomalijų detekcijos algoritmais galima identifikuoti triukšmą.

Projekte bus naudojamas atsitiktinių miškų algoritmu grįstas triukšmo valymas. HARF (angl. *high agreement random forest*). Metodas atsižvelgia į atskirų miško medžių nesutarimų skaičių prognozėse identifikuojant triukšmingus duomenis. Triukšmas bus traktuojamas pasinaudojus visų miške esančių medžių alijanso bendru sprendimu. Pavyzdžiui, jeigu nesutarimo rodiklis yra aukštas, nuo 70 – 90 proc, tai analizuojama duomenų eilutė yra laikoma kaip triukšmas, atvirkštiniu atveju traktuojama kaip švarūs duomenys [53].

Triukšmo šalinimas iš duomenų, padeda sukurti geresnį klasifikatorių. Atsitiktinių miškų atveju, išvalytuose duomenyse nebus sukuriama daug triukšmą atspindinčių medžių, kurie galimai blogai klasifikuos duomenis. Netgi išvalius ar ištrynus nedidelį procentą duomenų iš apmokymo duomenų imties, klasifikatoriaus prognozavimo kokybė gali stipriai pagerėti.

3. Tyrimų rezultatai ir aptarimas

Šioje darbo dalyje bus praktiškai pritaikomos 2 projekto dalyje išskirtos teksto vektorizavimo technikos kaip duomenų įvestis klasifikavimo algoritmams siekiant 1 projekto dalyje aprašytų analizių įvertinimo ant Lietuvos įmonių klientų atsiliepimų duomenų rinkinio. Teksto vektorizavimas bus atliekamas naudojantis komandine eilute ir Python programavimo kalba, o duomenų paruošimas, modeliavimas ir rezultato generavimas bus atliekamas R programavimo kalba.

3.1. Duomenys ir jų paruošimas

Pilnas lietuviškų atsiliepimų rinkinys buvo generuojamas sujungiant 2 skirtingus šaltinius: evertink.lt internetinį tinklalapį, kuriame pateikiami atsiliepimai apie internetines parduotuves ir socialiniame tinkle „Facebook“, atsiliepimų skiltyje parašytus įmonių klientų atsiliepimus. Evertink.lt tinklapyje parašyti atsiliepimai apima tik e-komerciją ir yra per siauri norint nagrinėti ne tik virtualius bet ir fizinius kanalus, kuriuose vartotojai sąveikauja su įmone. Dėl šios priežasties evertink.lt atsiliepimų rinkinys buvo papildytas „Facebook“ socialiniame tinkle parašytais vartotojų atsiliepimais, kur papildomai įtraukiamos ir tokios įmonės kaip pvz., degalinių tinklai, didieji prekybos centrai. Siekiant vieningo formato ir šių 2 šaltinių integracijos, buvo atliekamas papildomas turimų nestruktūrizuotų duomenų paruošimas. Siekiama, kad duomenų rinkinyje butų atsiliepimo indentifikavimo laukas (ID), šaltinis (evertink.lt ar Facebook), parduotuvės/paslaugų vykdytojo identifikatorius, įvertinimas $\in \{1, \ldots, 5\}$, atsiliepimo parašymo data (YYYY-MM-DD) ir atsiliepimo tekstas.

Internetinio tinklalapio evertink.lt duomenys buvo gauti jau surinkti, todėl duomenų ištraukimo iš tinklalapio žingsnis nebuvo vykdomas. Pirmiausia, šio šaltinio visų turimų parduotuvių failai .rds formate, sujungiami. Toliau, paliekami tik reikalingi kintamieji galutiniam formatui. Įvertinimo skalėje pateiktas įvertinimas svyruoja nuo 1 iki 100, šis laukas papildomai modifikuojamas: skalė sumažinama iki dešimtbalės (padalinama iš 10 ir gautas skaičius apvalinamas pagal matematines taisykles), toliau iki penkiabalės, kur 1, 2 įvertinimams priskiriamas 1 ir t. t. Ištrinami nelietuviški atsilepimai, pasinaudojus automatine kalbos detekcijos funkcija, taip pat pašalinami atsiliepimai parašyti vartotojų, kurie save susiejo su asmenine „Facebook“ paskyra. Daroma prielaida, kad vartotojas parašęs atsiliepimą evertink.lt galėjo ta patį parašyti ir įmonės socialinio tinklo paskyroje, taip bus išvengiama galimp nuomonės dubliavimosi.

Socialinio tinklo „Facebook“ įmonių atsiliepimų duomenys buvo surinkti autorės, pasinaudojus įrankiu „Parsehub“. Duomenys buvo ištraukti iš verslo paskyrų, kuriose buvo aktyvi atsiliepimų skiltis. Iš turimų .json formato failų, buvo sugeneruojami kintamieji reikalingam formatui. Įvertinimo lauko skalė svyravo nuo 1 iki 5 senesniuose atsiliepimuose, o naujesniuose atsiliepimo poliariškumas buvo klasifikuojamas į ‘rekomenduoja’ ir ‘nerekomenduoja’. Šiuo atveju, ‘rekomenduoja’ požymiui buvo priskiriamas įvertinimas 5, o ‘nerekomenduoja’ - 1. Taip pat, kaip ir evertink.lt rinkinyje, pašalinami nelietuviški atsiliepimai.

Sutvarkyti duomenų šaltiniai, sujungiami į projekte naudojamą duomenų rinkinį. Bendrai rinkinį sudaro 19270 atsiliepimai. Eilės tvarka, visoms duomenų eilutėms priskiriamas unikalus ID, dėl tolimesnio atsekamumo.

3.2. Klasių žymos priskyrimas

Siekiant suklasifikuoti atsiliepimus kuo tiksliau, nuspręsta klases priskirti rankiniu būdų. Rankinis klasių priskyrimas priklauso nuo žyminčio asmens subjektyvumo. Pirmiausia, norint išvengti visų atsiliepimų peržiūrėjimo, įvertinimai, kurio skaitinė reikšmė yra 1, priskiriama klasė 1 (projekto tikslinis objektas), o įvertinimams, kurių skaitinė reikšmė 5, priskiriama klasės žyma 0. Tai logiškai paaiškinama tuo, kada vartotojas kuris yra labai nepatenkintas nevertins įmonės geriau nei 1, o klientai, kurių patirtis sąveikaujant su įmone yra itin gera, nevertins įmonės žemiau 5. Taip išvengiama nereikšmingo laiko resursų panaudojimo.

Tolimesnis darbas tęsiamas su įvertinimais, kurių skaitinės reikšmės 2, 3 arba 4. Pateikiamas tokių atsiliepimų pavyzdys (kalba netaisyta), kur vartotojų įvertinimas skiriasi nuo atsiliepimo konteksto, t. y. įvertinimas žemas, tačiau kontekstas teigiamas ir atvirkščiai:

2 lentelė. Atsiliepimų pavyzdžiai

Įmonės pavadinimas	**Atsiliepimas**	**Įvertinimas**
knygos.lt	Ilgokai reikia laukti prekių.	4
pigu.lt	reikejo penkis kartus skambinti kol atveze suduzusias lekstes	4
supakuota.lt	užsisakiau rudas dėžutes, o gavau baltas	4
„Antis“ restoranas	Maisto laukeme valanda. Salotos su vistiena patiektos belekaip. Salotu lapai sumesti didziausiais gabalais. Vistienos gabalas irgi didziausias. Truputi nusivylem, bet maistas visai skanus.	3
Ermitažas	Šiauliuose tilžei puikus aptarnavimas.	3
balduturgus.lt	MAN ASMENISKAI VISI BALDAI TIESIOG PUIKUS	3
Girstučio baseinas	tevams leidziantiems savo atzalas i plaukimo treniruotes rekomenduoju turint galimybe vis atvykti pastebeti treneriu darbo.	2

Priskiriant klasių žymas, buvo mąstoma iš įmonės pozicijos, kokį atsiliepimo kontekstą būtų aktualu identifikuoti ir būtų gaunama pridėtinė informacija. Buvo vadovaujamasi principu, kad jeigu atsiliepime yra neigiamo konteksto, tada priskiriama klasės žyma 1. t. y. netgi tokie atsilepimai kaip ‘*viskas gerai, tik galėtų būti siuntimas greitesnis*‘ buvo suklasifikuojami kaip klasė 1, nes iš tokio atsiliepimo galima identifikuoti, kad įmonės probleminė sritis – prekių transportavimas. Vadovaujantis šia logika, rankiniu būdų klasės žyma buvo priskirtas 2968 atsiliepimams.

Visgi, toks rankinis žymų priskyrimas kvestionuotinas dėl įvertinimo reikšmių 1 ir 5, taip pat, galbūt buvo blogai suklasifikuoti ir rankiniu būtų tikrinti atsiliepimai. Dėl šios priežasties nuspręsta atlikti pakartotinį rankinį klasių žymų patikrinimą. Tam bus sukuriamas mašininio mokymosi modelis, tikslui identifikuoti atsiliepimus prie EER slenksčio. Šiame EER taške bus gaunami atsiliepimai, dėl kurių klasės atsitiktinių miškų algoritmas yra neapsisprendęs t. y. miško medžiai suprognozuoja tikimybę, kad atsiliepimo klasė yra 1 ir 0 beveik vienodomis proporcijomis. Taip bus gaunami sudėtingi klasifikavimo atvejai.

Pradžioje visas duomenų rinkinys vektorizuojamas LSI metodu, pakeičiant tekstinį lauką į jį reprezentuojantį 128 požymių vektorių. Gautas duomenų rinkinys išskaidomas į 2 lygius duomenų rinkinius išlaikant identišką klasių balansą. Pasinaudojus atsitiktinių miškų algoritmu, pirmiausiai

modeliuojama su pirmojo rinkinio duomenimis, o prognozė išskaičiuojama ant antrojo rinkinio. Išskaičiuojamas optimalus taškas, kuriame turime EER slenkstį. Pirmojo bandymo atveju toks taškas yra 0.2960. Pasirenkama toleruotina riba – 0.1 į abi puses, t. y. pakartotinai patikrinimui bus gaunami visi atsiliepimai kurie svyruoja (0.1960, 0.3960) intervale. Toliau, identiška procedūra atliekama apkeitus duomenų rinkinius vietomis. Antruoju bandymu optimalus taškas yra 0.3180. Iš viso, pakartotinam patikrinimui buvo išrinkti 2805 atvejai. Po patikrinimo, buvo ištrinti nereikšmingi atsiliepimai (108 atvejai) (žr. lentelė 3)

3 lentelė. Ištrintų atsiliepimų pavyzdžiai

Įmonės pavadinimas	**Atsiliepimas**
Novaturas	Kelione i Kreta Agapi beach viesbutis, isvykimas 05 28 grizimas 06 04 ,viskas iskaiciuota, sios dienos kaina 1500e, parduodu uz 1000e. Dviems. SKUBIAI
pigu.lt	Žavingos merginos ;)
Lemon gym	Perleidziu multi naryste!!! Galioja iki 2017-05-05

Matoma, kad tokie atsiliepimai jokios pridėtinės vertes kuriant modelius – neduos, nes yra nereikšmingi, todėl modelis gali nesugebėti jų suklasifikuoti tinkamai. Taip pat, 61 atsiliepimas, kuris turėjo įvertinimą 5, buvo perklasifikuotas į 1 klasę, nes turėjo neigiamo konteksto. Po pakartotinio patikrinimo, klasės žymos pasikeitė 12 proc. atsiliepimų.

3.3. Žvalgomoji analizė

Galutiniam duomenų rinkiniui gauti bei rinkinio kokybei pagerinti, buvo pašalinti nereikšmingi atsiliepimai. Triukšmo šalinimui panaudotas HARF metodas. Metodas buvo pritaikytas panaudojus 500 sprendimų medžius, o nustatytas sutarimo rodiklis siekė 0.75. Po triukšmo šalinimo, buvo ištrinti 622 atsiliepimai, o tai sudarė 3.25 proc. viso rinkinio.

4 lentelė. Atsiliepimų triukšmo pavyzdžiai

Įmonės pavadinimas	**Atsiliepimas**
Girstučio baseinas	Viskas labai gerai ir smagu, tik paskutiniu metu labai šalta moterų rūbinėje ir praėjime link jos nuo baseino pusės.
Lidl Lietuva	I live in the Uk and it's the best shop ever!Džiugu kad Lidl atsiranda ir Lietuvoje! Kuo daugiau jūsų, tuo geriau
Robotikos Akademija	Sūnaus laukiamiausia diena savaitėje yra šeštadienis, nes tada jis eina į robotikos būrelį. Nerealiai blizga akys pasakojant ką veikė būrelyje ir grįžęs jau laukia kito savaitgalio. Vienintelė paskata būrelis, gali sakyti, kad negaus planšetės ar telefono, tai nėra taip baisu nei kai pasakai, kad tai bus tavo paskutinis kartas robotikos būrelyje, jei elgsies netinkamai :) Didžiausia motyvacija šis būrelis!

Iš lentelės 4 matyti, kad triukšmu buvo laikomi atsiliepimai kuriuose buvo panaši proporcija teigiamo ir neigiamo konteksto (1 atsiliepimas ir 4 lentelės) arba tokie atsiliepimai, kurie yra abstraktūs.

Toliau bus dirbama su duomenų rinkiniu, kuriame nėra triukšmo, bendrai rinkinį sudaro 18539 atsiliepimai. Klasių balansas rinkinyje:

5 lentelė. Klasių balansas duomenų rinkinyje

Klasė	**Atsiliepimų skaičius**	**Dalis, %**
0 (teigiamas poliariškumas)	13846	75 proc.
1 (neutralus / neigiamas poliariškumas)	4693	25 proc.

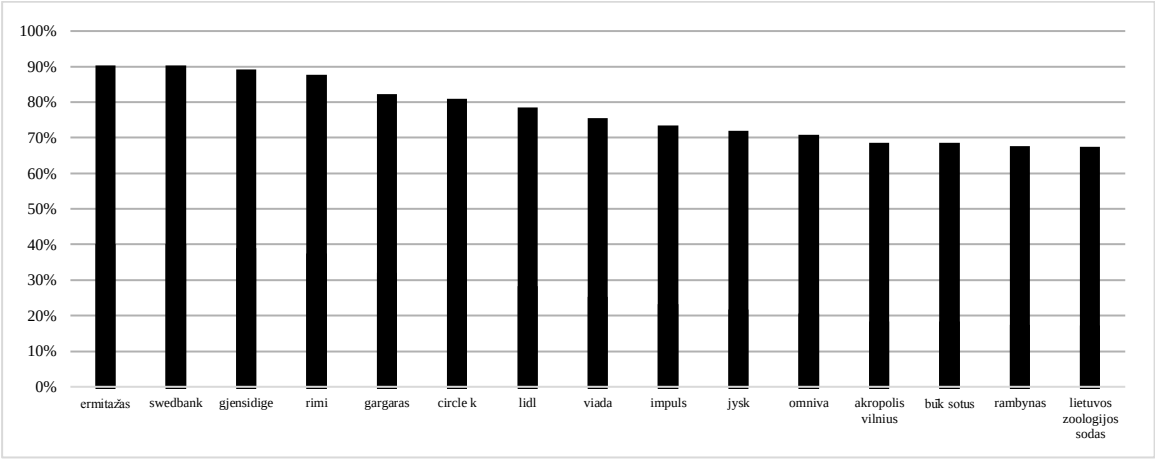
Nors rinkinyje teigiamo konteksto atsiliepimų daugiau, tačiau klasių disbalanso dar nėra, todėl papildomų modifikacijų pagerinti duomenų rinkinį nebus daroma. Duomenų rinkinį sudaro evertink.lt ir Facebook šaltinių duomenys, kuriame yra 134 unikalių įmonių atsiliepimai, pvz., bankas „Swedbank“, užkandinė „Keule Ruke“ ar degalinė „Circle K“. Didelis įvairių sričių įmonių kontingentas, leis sukurti universalesnį klasifikatorių, kuris gebės aptikti skirtingas verslo problemas. Atitinkamai rinkinyje yra 66 proc. evertink.lt tinklalapio atsiliepimų ir 34 proc. „Facebook“ įmonių atsiliepimų skiltyje parašytų atsiliepimų. Daugiausiai atsiliepimų rinkinyje turi šios 3 įmonės per skirtingus šaltinius:

6 lentelė. Daugiausiai atsiliepimų turinčios įmonės

Šaltinis	**Įmonė**	**Atsiliepimų skaičius**
Facebook	Lidl	360
Facebook	drabuzelis.lt	255
Facebook	groziobaze.lt	251
Evertink.lt	pigu.lt	3431
Evertink.lt	knygos.lt	1698
Evertink.lt	neriba.lt	1610

Pagrindinės evertink.lt šaltinyje esančios 3 įmonės sudaro netgi 55 proc. visų šaltinio atsiliepimų. Facebook šaltinio duomenys yra universalesni, apima daugiau skirtingų verslo sričių ir įmonių, o 3 daugiausiai atsiliepimų turinčios įmonės sudaro tik 8 proc. visų šaltinio atsiliepimų.

Toliau pateikiamas įmonių sąrašas, kurios turi daugiausiai neigiamų atsiliepimų įvertinant bendrą įmonės atsiliepimų skaičių ir atmetant tas, kurios turi mažiau nei 50 atsiliepimų (žr. 6 lentelė)



13 pav. Didžiausią neigiamų atsiliepimų santykį turinčios įmonės

Šių 15 įmonių atsiliepimai buvo gauti iš Facebook socialinio tinklo. Pastebima, kad Facebook socialiniame tinkle, vartotojai linkę labiau rašyti neigiamus atsiliepimus, nei evertink.lt tinklalapyje.

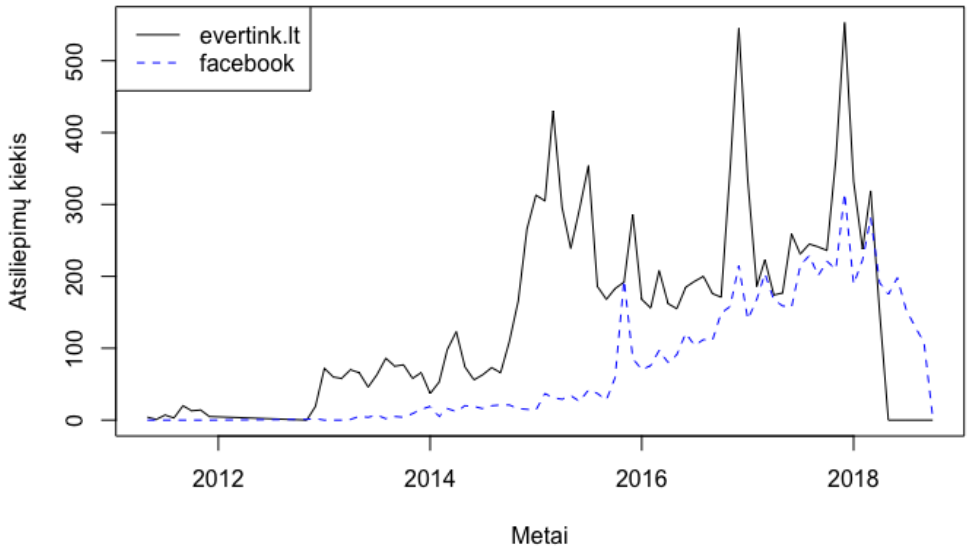
Iš visų 134 įmonių, 43 įmonių neigiamų atsiliepimų skaičius buvo didesnis nei teigiamų. Įmonėms tai ypač informatyvi informacija, kuri padėtų identifikuoti problemines sritis, taip pat, patvirtinama ir prielaida, kad socialiniai tinklai ypač populiarūs nuomonės sklaidai.

Svarbu identifikuoti kokios pagrindinės priežastys veikia neigiamą patirtį vartotojui, kas tą sąlygoja. Projekto tiksliniam objektui – klasei 1, nubraižomas žodžių debesis, kuris išskiria pagrindines atsiliepimų tematikas (žr. 14 pav.)

14 pav. Neigiamą kontekstą atvaizduojantis žodžių debesis

Žodžių debesis išskiria tokias sritis, dėl kurių vartotojai skundžiasi, iš pav. 14 matoma, kad tokios sritys yra: aptarnavimas, personalo komunikacija, prekių ar paslaugų kokybė, pristatymas, kaina, vėluojančios siuntos ir kt. Tokia skirtinga sričių variacija atsiliepimuose, leis kuriamam klasifikatoriui aptikti plataus spektro problemines sritis.

Šaltinio evertink.lt atsiliepimai buvo surinkti iki 2018 m. gegužės mėnesio, Facebook šaltinio iki 2018 m. lapkričio mėnesio. Pateikiamas grafikas, parodantis atsiliepimų skaičiaus kitimo priklausomybę nuo metų (žr. 15 pav.)



15 pav. Atsiliepimų skaičiaus kitimo priklausomybė nuo metų

Grafike pateikta dinamika leidžia pamatyti, kad Facebook socialiniame tinkle parašytų atsiliepimų skaičius turi tendenciją didėti. Per pastaruosius 4-5 metus, šis socialinis tinklas pasidarė populiari erdvė išreikšti nuomonei. Šiame tinkle yra pastebimas vienas didžiausių naujų vartotojų augimas. Tuo pačiu, verslai reagavo į pasaulines tendencijas ir susikūrė savo įmonių paskyras. Dalis verslų, suprasdami vartotojų patirties identifikavimo svarbą, paskyrose įgalino atsiliepimų skiltį. Kitame šaltinyje, evertink.lt, pastebimas žymus atsiliepimų skaičiaus padidėjimas atitinkamais periodais. Dauguma šių padidėjimų yra paskutinį arba pirmą metų mėnesį, kada išauga vartotojų perkamumas dėl sezoniškumo (pvz., švenčių).

Sentimento poliariškumo nustatymui pritaikomi leksikonu grįsti metodai, kurių pasiektas tikslumas bus atsvaros tašku vertinant mašininiu mokymusi pasiekto tikslumo kokybę. Bus išbandomi 2 žodynų tipai – tai statinis, kur Lietuvių kalbai teigiamo ir neigiamo konteksto žodžius suformavo Chen‘as ir kt. (2015), bei dinaminis, kuris bus sugeneruojamas iš turimo duomenų rinkinio. Atlikus šią procedūrą, gautas tikslumas dinamišku žodynu siekė 81 proc. Statinis žodynu, kai sentimento poliariškumas buvo vertinamas pagal jau pateiktus poliariškumą nusakančius rinkinius, tikslumo vidurkis buvo mažesnis už dinamišku žodynu pasiektą ir siekė 72 proc., t. y. šis metodas blogiau prognozuoja nei atsitiktinis spėjimas. Atsitiktiniu spėjimu potencialus tikslumas - 75 proc., kuris gaunamas visus duomenų rinkinio stebėjimus klasifikuojant kaip 0 klasę. Toliau, didžiausias pasiektas tikslumas leksikonu grįstais metodais bus kaip informacinis vertinant mašininiu mokymusi sukurtų kombinacijų tikslumą, bei bus siekiama pagerinti 81 proc. sentimento poliariškumo prognozavimo tikslumą.

3.4. Sentimento detekcijos modeliai

Sentimentų detekcijos uždavinio geriausiam modeliui surasti bus palyginami 8 vektorizavimo metodai: Doc2Vec_dbow, Doc2Vec_dbow_w, Doc2Vec_dm_m, sent2vec (FastText), LSI, LDA, RP ir BERT, bei 4 mašininio mokymosi modeliai RP, SVM, LogR ir XGBoost. Tuo pačiu, norint patikrinti ar vektorių dimensionalumas turi įtakos modeliavimo rezultatams, bus išbandomas 3 skirtingų dimensionalumų vektorizavimas, kur dimensijų dydis atitinkamai bus 64, 128 ir 264. Galutiniame variante bus išbandyta 96 (V x M x D) skirtingi variantai iš kurių bus pasirinktos 3 tiksliausiai gebančios prognozuoti konstrukcijos: po 1 kiekvieno dimensionalumo atveju.

Mašininio modelio geriausių parametrų nustatymui bus įdiegiamos skirtingos parametrų variacijos. Modelių validacijai naudojamas kryžminio patikrinimo metodas, kuris dėl didelio laiko resursų naudojimo modeliuojant, bus apribojamas iki 2 duomenų rinkinio dalinimų (angl. *folds*). Siekiant išlaikyti lygią klasių proporciją duomenų rinkinio daliniuose, bus įgalinamas stratifikuotas kryžminio patikrinimo metodas.

Geriausias vektorizavimo tipo modelis bus pasirinktas pagal Kappa tikslumo matą, kadangi klasių proporcijos nėra lygios ir teigiamo poliariškumo atsiliepimų yra daugiau. Toliau, geriausio vektorizavimo tipo ir modelio kombinacijos bus lyginamos tarpusavyje pasinaudojus ROC, DET ir PR kreivėmis, taip bus identifikuojama geriausiai prognozuojanti kombinacija.

3.4.1. 64D duomenų rinkinys

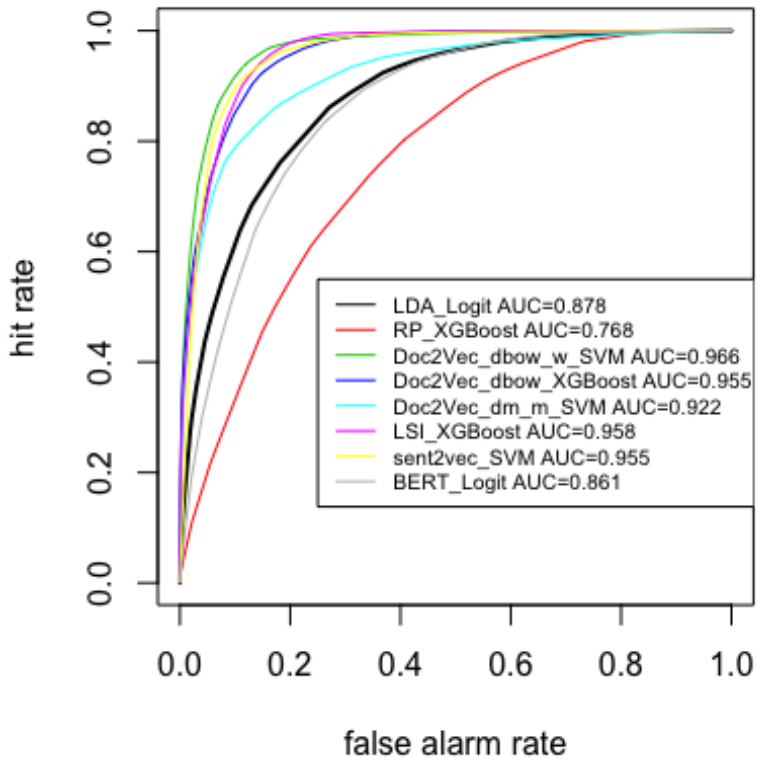
Pradedama nuo atsiliepimo tekstinio lauko išskaidymo į 64 dimensijų vektorių, kur kiekviena eilutė reprezentuos atsiliepimą. Prie vektoriaus prijungiama sentimento klasė. Gaunamas duomenų rinkinys

– matrica, kurios dydis 18539 x 65, kuri tinkama kaip įvestis į pasirinktus mašininio mokymosi modelius. Vektorizavimo tipų sukūrimas ir modelių mokymas užtruko apie 10 val. Pateikiama gauta Kappa ir AUC tikslumo įverčių lentelė (žr. 7 lentelė)

7 lentelė. Tikslumo matų Kappa ir AUC rezultatai skirtingoms 64D duomenų rinkinio kombinacijoms

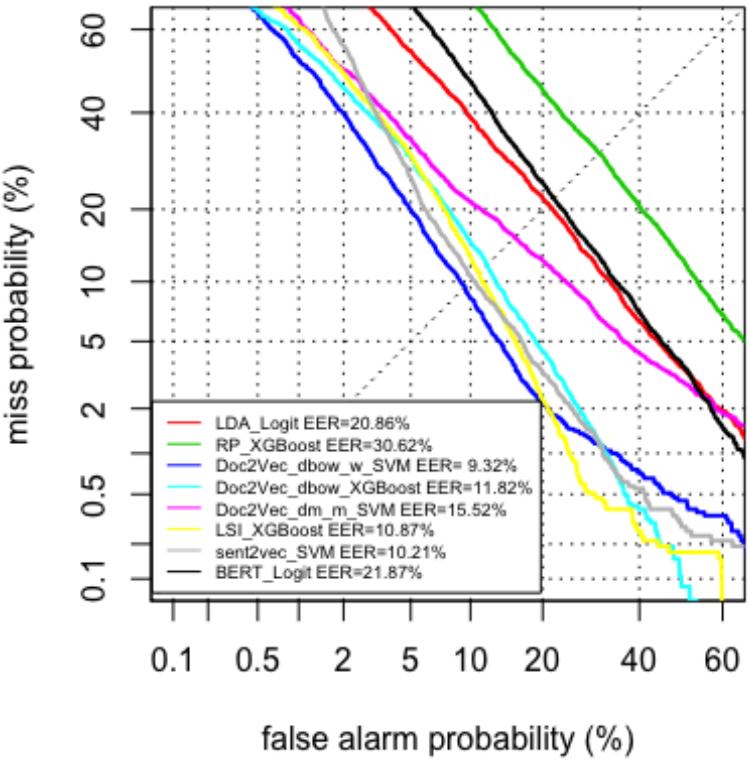
	SVM-Lin		**RF**		**LogR**		**XGBoost**	
	Kappa	AUC	Kappa	AUC	Kappa	AUC	Kappa	AUC
Doc2Vec_dbow_w	**0,78**	**0,97**	0,68	0,96	0,77	0,97	0,77	0,97
Doc2Vec_dbow	0,70	0,95	0,66	0,95	0,69	0,95	**0,71**	**0,96**
Doc2Vec_dm_m	**0,68**	**0,92**	0,57	0,93	0,63	0,92	0,66	0,94
LDA	0,51	0,88	0,50	0,87	**0,51**	**0,88**	0,51	0,88
LSI	0,71	0,95	0,72	0,96	0,71	0,95	**0,73**	**0,96**
sent2vec	**0,75**	**0,96**	0,69	0,95	0,74	0,96	0,72	0,96
BERT	0,50	0,86	0,43	0,84	**0,49**	**0,86**	0,49	0,87
RP	0,29	0,73	0,28	0,76	0,31	0,75	**0,32**	**0,77**

Iš lentelės matoma, kad atsitiktinių miškų algoritmais beveik visais atvejais prognozuodavo blogiausiai, likusių metodų prognozavimo tikslumo kokybė panaši. Blogiausiai atsiliepimus visais atvejais reprezentavo atsitiktinio indeksavimo (RP) vektorizavimo metodas, šio metodo tikslumas žemas, tačiau metodas vis tiek geriau prognozuodavo nei atsitiktiniai spėjimai (AUC > 0.5). Papildomai pateikiamos kombinacijų tikslumą įvertinančios kreivės, siekiant išrinkti geriausią kombinaciją.



16 pav. Geriausių 64D kombinacijų ROC kreivių grafikas

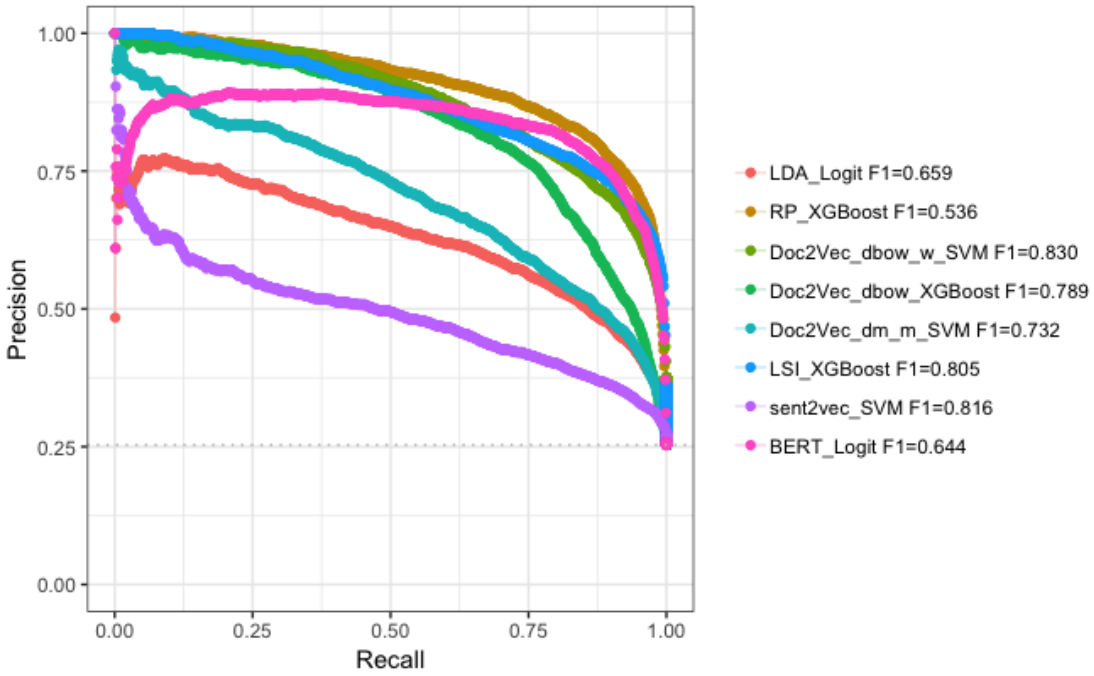
Paveiksle 16, nubraižytos geriausių kombinacijų ROC kreivės. Iš grafiko pateikiamas ir kombinacijos AUC matas. Arčiausiai kairiojo viršutinio taško yra Doc2Vec_dbow_w vektorizavimo būdo ir SVM modelio kombinacija. Šis grafikas atvaizduoja, kad kombinacija yra tiksliausia, tas pats buvo identifikuota ir ankščiau pateiktoje lentelėje (žr. 7 lentelė). Iš šios kombinacijos AUC įverčio (0.966) galima teigti, kad kombinacija prognozuoja ypač arti idealaus atvejo. Kombinacija – RP ir XGBoost yra atvejis, kur iš ROC kreivės formos ir pozicijos matomas netikslaus klasifikavimo pavyzdys.



17 pav. Geriausių 64D kombinacijų DET kreivių grafikas

Kitame grafike – DET, pateikiamos klaidų galimybės, kur EER galima interpretuoti kaip vienodą klaidingo priėmimo ir klaidingo atmetimo dažnumą išreikštą procentais. Prie žemiausio EER yra jau ankščiau, iš AUC mato, identifikuota kombinacija. Matoma, kad surasta geriausia kombinacija kreivės pabaigoje nukrypsta ir prognozavimas pablogėja, šioje vietoje konkrečiai pablogėja 0 klasės prognozavimas.

Iš toliau pateikiamo PR grafiko, galima pamatyti teisingai klasifikuoto projekto tikslinės klasės santykį su bendra klasės proporcija. Pridedamas ir F-mato įvertis, korektiškesniam įvertinimui.



18 pav. Geriausių 64D kombinacijų PR grafikas

Kreivės nėra tokios informatyvios kaip F-matas, kadangi yra viena ant kitos ir persidengia. Iš grafiko patvirtinama, kad tiriamame dimensionalume geriausia klasifikavimo kombinacija yra gauta vektorizuojant atsiliepimus Doc2Vec_dbow_w metodu ir sukuriant klasifikatorių tiesiniu SVM modeliu.

3.4.2. 128D duomenų rinkinys

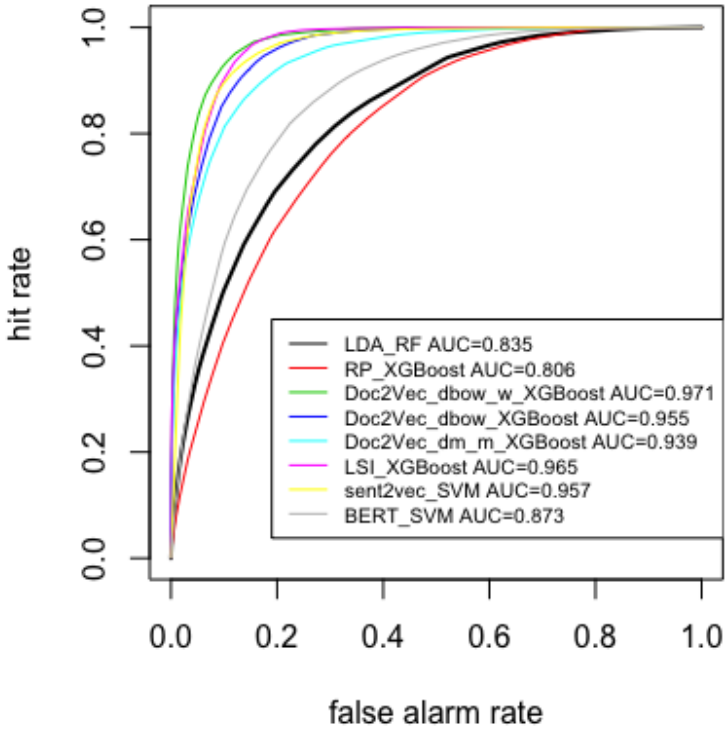
Toliau, tiksliausiai prognozuojančios konstrukcijos paieška tęsiama atsiliepimą reprezentuojant 128 dimensijų vektoriumi. Atliekant identišką procedūrą, gaunama matrica, kurios dydis 18539 x 129. Vektorizavimo tipų sukūrimas ir modelių mokymas užtruko apie 19 val., šių kombinacijų įgalinimas sunaudojo daugiau laiko resursų nei ankstesnysis bandymas. Pateikiama gauta Kappa ir AUC tikslumo įverčių lentelė (žr. 8 lentelė)

8 lentelė. Tikslumo matų Kappa ir AUC rezultatai skirtingoms 128D duomenų rinkinio kombinacijoms

	SVM-Lin		**RF**		**LogR**		**XGBoost**	
	Kappa	AUC	Kappa	AUC	Kappa	AUC	Kappa	AUC
Doc2Vec_dbow_w	0,78	0,97	0,69	0,96	0,77	0,97	**0,78**	**0,97**
Doc2Vec_dbow	0,71	0,95	0,66	0,95	0,70	0,95	**0,71**	**0,96**
Doc2Vec_dm_m	0,67	0,92	0,57	0,93	0,62	0,92	**0,67**	**0,94**
LDA	0,41	0,81	**0,44**	**0,84**	0,40	0,81	0,43	0,84
LSI	0,74	0,96	0,73	0,96	0,74	0,96	**0,75**	**0,97**
sent2vec	**0,75**	**0,96**	0,70	0,95	0,74	0,96	0,73	0,96
BERT	**0,53**	**0,87**	0,43	0,85	0,52	0,88	0,52	0,88
RP	0,36	0,78	0,30	0,79	0,35	0,79	**0,39**	**0,81**

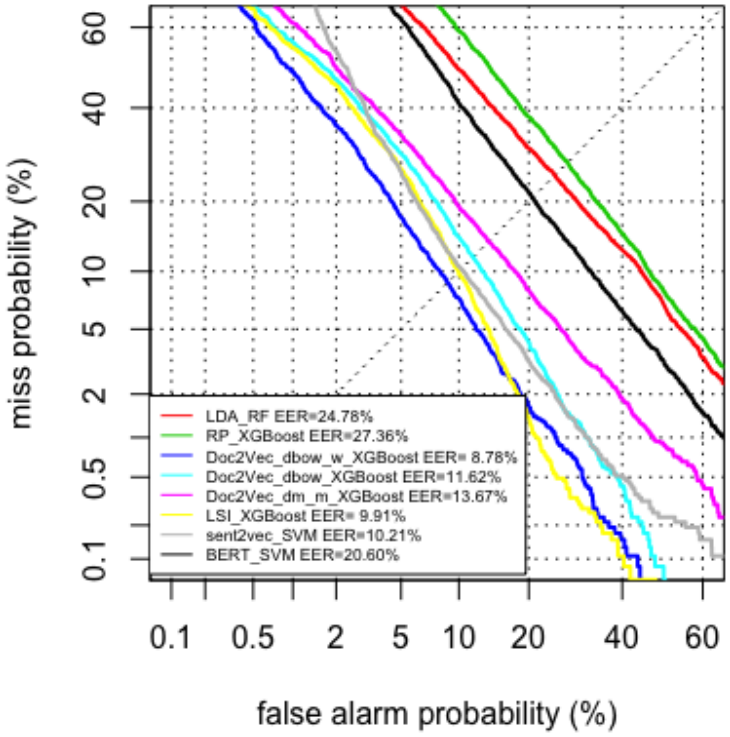
Pastebima, kad XGBoost algoritmas daugumoje atveju buvo tiksliausias, kadangi padidėjus modeliuojamų požymių skaičiui, šis algoritmas gebėjo geriau apdoroti didesnius duomenų rinkinius.

Didesnis dimensijų skaičius bendrai padidino daugumos vektorizavimo būdų ir modelių kombinacijų prognozavimo tikslumą, tačiau padidėjimas nėra žymus. Pateikiamas geriausių kombinacijų ROC kreivių grafikas (žr. 19 pav.)



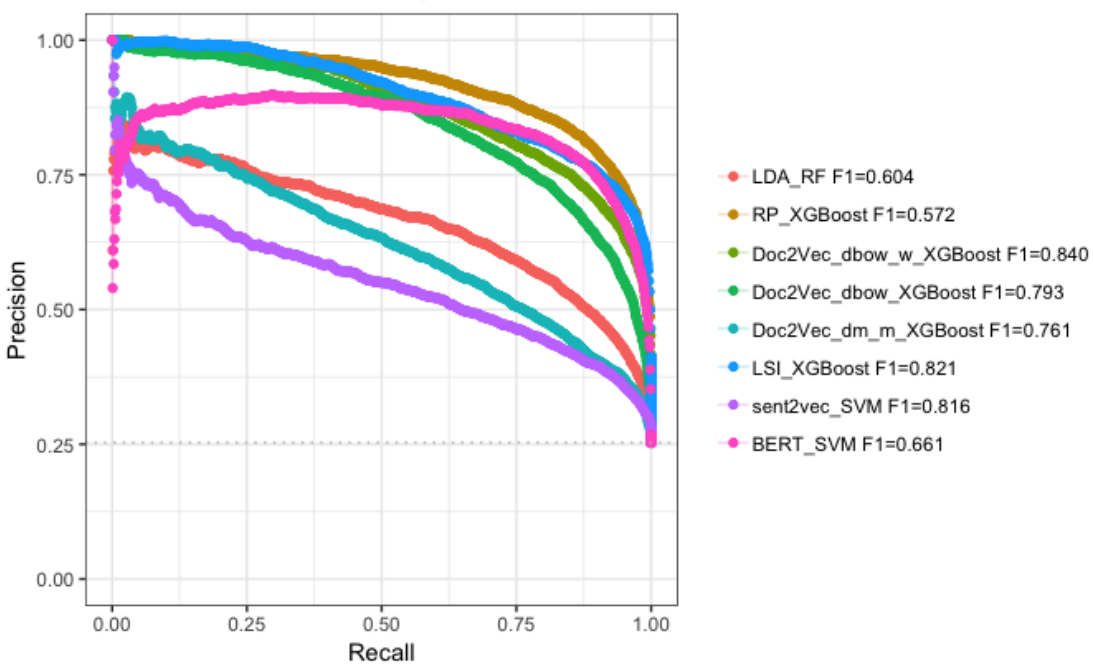
19 pav. Geriausių 128D kombinacijų ROC kreivių grafikas

ROC kreivių grafike matomas kreivių pasikeitimas, kuris parodo, kad beveik visų kombinacijų prognozavimo tikslumas yra pagerėjęs, tačiau LDA vektorizavimo būdo geriausios kombinacijos prognozavimas – suprastėjo.



20 pav. Geriausių 128D kombinacijų DET kreivių grafikas

Iš DET kreivės matoma tokia pati tendencija, kombinacijų klydimo procentinė dalis – mažesnė. Geriausiai klasifikuojantis vektorizavimo būdas išliko Doc2Vec_dbow_w, tačiau pasikeitė klasifikatoriaus algoritmas, nagrinėjamo dimensionalumo atveju, geriausias klasifikatorius buvo gautas XGBoost algoritmu. Toliau pateikiamas PR kreivių grafikas, kuris bus palyginamas su ankstesniu atveju (žr. 18 pav.) ir įvertinamas prognozuojamos klasės prognozavimo pokytis (žr. 21 pav.)



21 pav. Geriausių 128D kombinacijų PR grafikas

Prognozavimo pagerėjimas taip pat identifikuojamas ir iš F-mato įverčių. Vektorizavimo tipo sent2vec F-matas nepasikeitė lyginant su praėjusiu atveju, o LDA pablogėjo. Iš visų pateiktų kreivių – DET, ROC ir PR, tiksliausia klasifikavimo kombinacija gauta Doc2Vec_dbow_w vektorizavimo būdu ir XGBoost klasifikavimo algoritmu.

3.4.3. 256D duomenų rinkinys

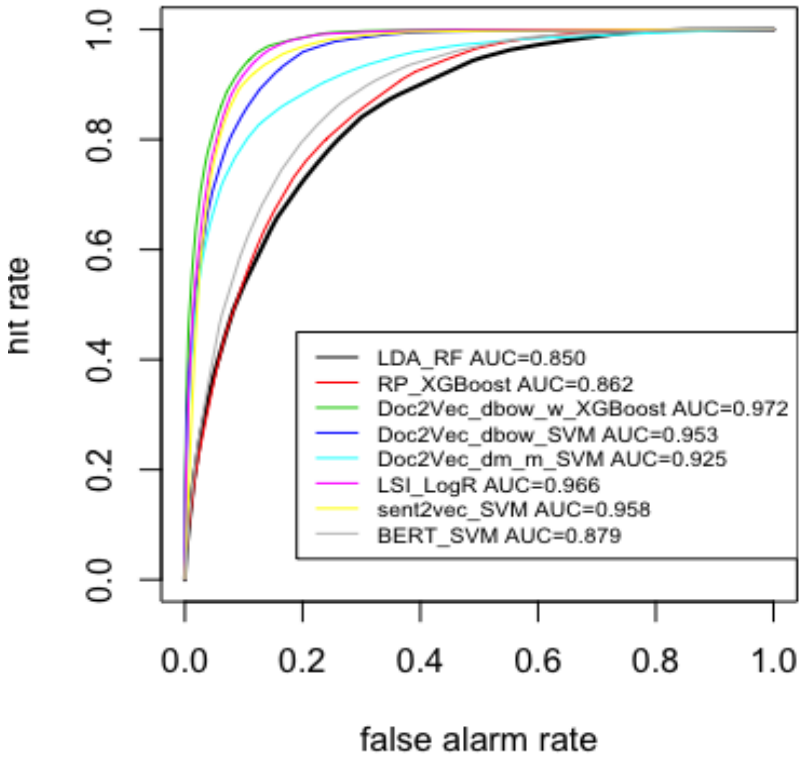
Paskutinis bandymas atliekamas atsiliepimą pakeičiant į jį reprezentuojantį 256 dimensijų vektorių. Turima matrica, kurios dydis 18539 x 257. Vektorizavimo tipų sukūrimas ir modelių mokymas užtruko apie 42 val., kuo didesnis vektoriaus dimensijų skaičiaus, tuo proporcingai išilgėjo visų kombinacijų sukūrimas ir modeliavimas. Pateikiama naujai gauta Kappa ir AUC tikslumo įverčių lentelė (žr. 9 lentelė).

9 lentelė. Tikslumo matų Kappa ir AUC rezultatai skirtingoms 256D duomenų rinkinio kombinacijoms

	SVM-Lin		**RF**		**LogR**		**XGBoost**	
	Kappa	AUC	Kappa	AUC	Kappa	AUC	Kappa	AUC
Doc2Vec_dbow_w	0,78	0,97	0,70	0,96	0,77	0,97	**0,78**	**0,97**
Doc2Vec_dbow	**0,72**	**0,95**	0,64	0,95	0,70	0,95	0,71	0,96
Doc2Vec_dm_m	**0,68**	**0,93**	0,56	0,93	0,62	0,92	0,67	0,94
LDA	0,44	0,83	**0,46**	**0,85**	0,43	0,83	0,46	0,85
LSI	0,76	0,97	0,73	0,96	**0,77**	**0,97**	0,76	0,97

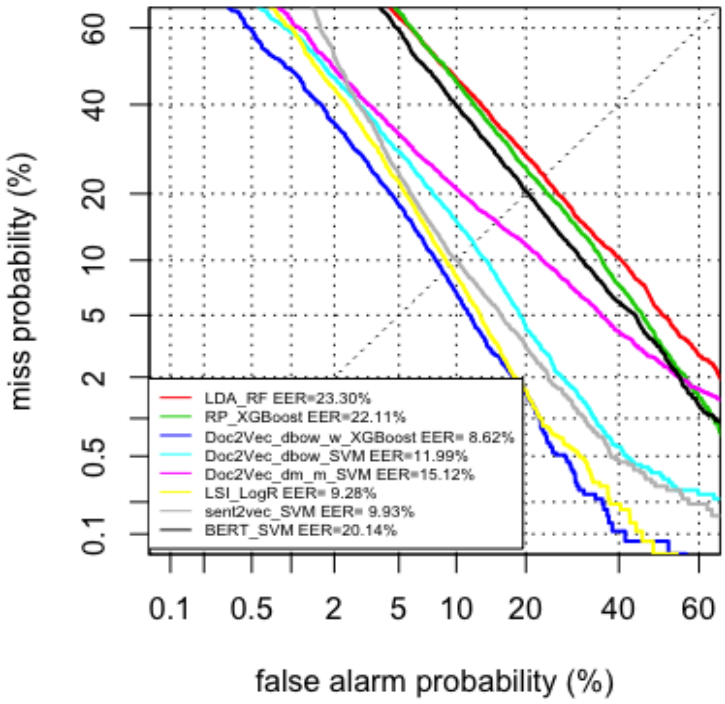
	SVM-Lin		RF		LogR		XGBoost	
	Kappa	AUC	Kappa	AUC	Kappa	AUC	Kappa	AUC
sent2vec	**0,76**	**0,96**	0,70	0,95	0,75	0,96	0,73	0,96
BERT	**0,54**	**0,88**	0,41	0,84	0,53	0,88	0,53	0,88
RP	0,48	0,85	0,35	0,83	0,47	0,85	**0,49**	**0,86**

Dimensionalumo padidinimas davė teigiamų rezultatų gerinant prognozavimo tikslumą. Tikslumo padidėjimas, taip pat kaip ir ankstesniu atveju, nėra žymus. Įžvelgiamas ir kai kurių kombinacijų tiklsumo sumažėjimas, pvz., Doc2Vec_dbow vektorizavimo ir XGBoost algoritmo kombinacijoje, tačiau nagrinėjamo dimensionalumo atveju, Doc2Vec_dbow vektorizavimo būdu geresnį rezultatą pateikė SVM modelis, kurio prognozavimas geresnis nei praėjusio atvejo geriausios šio vektorizavimo būdo kombinacijos. Didžiausias tikslumo pagerėjimas pastebimas RP vektorizavimo metode, tačiau šis metodas visais atvejais prognozuodavo blogiausiai, todėl toks didelis pagerėjimas reikšmingos naudos neduoda.



22 pav. Geriausių 256D kombinacijų ROC kreivių grafikas

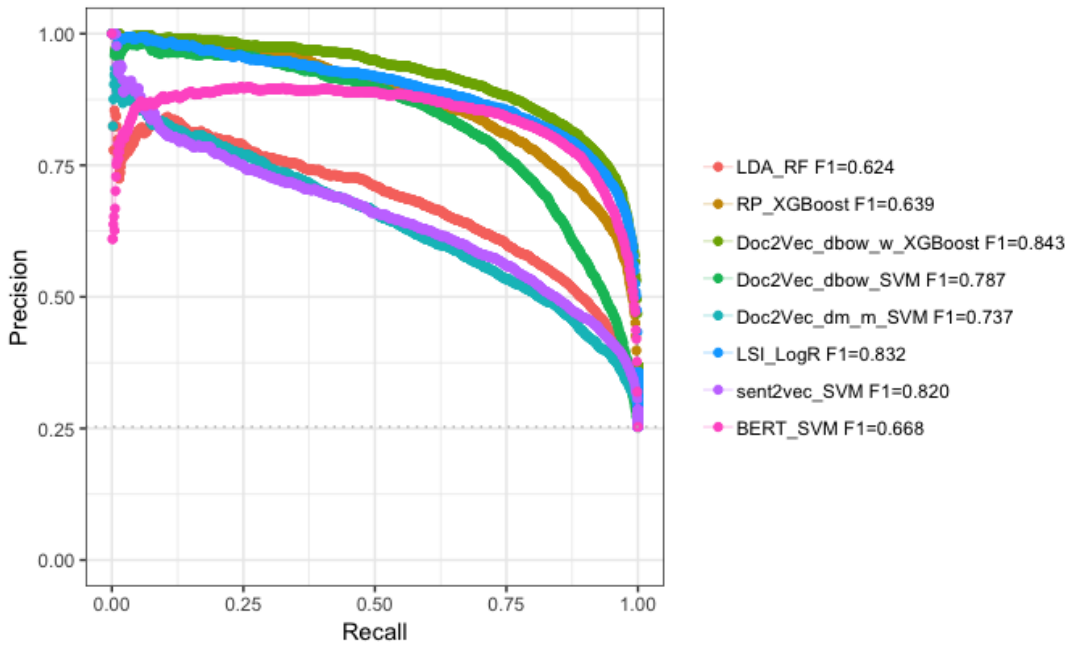
Iš ROC kreivė matomas jau ankščiau pastebėtas rezultatas, geriausiai klasifikuojanti kombinacija išlieka ta pati.



23 pav. Geriausių 256D kombinacijų DET kreivių grafikas

Tiksliausiai prognozuojanti kombinacija Doc2Vec_dbow_w ir XGBoost, kurios EER= 8.62%, t. .y. šis metodas blogai klasifikuoja tik daugiau nei 8 proc. duomenų rinkinio. Iš 3 geriausiai klasifikuojančių vektorizavimo metodų ir modelių kombinacijų, LSI ir LogR kombinacijoje įvyko didžiausias tikslumo progresas, Kappa matas padidėjo apytiksliai per 0.02 punktus, kas rodo, kad šis metodas imlesnis didesnėms dimensijoms ir turi perspektyvų pagerinti Doc2Vec_dbow_w vektorizavimo būdo prognozavimo tikslumą, įvedant dar didesnį dimensijų skaičių. Kitų kombinacijų padidėjimas nėra žymus ir panašu, dimensionalumo didinimas reikšmingo tikslumo padidėjimo neduoda.

Pateikiamas ir PR kreivų grafikas su F-mato įverčiais.



24 pav. Geriausių 256D kombinacijų PR grafikas

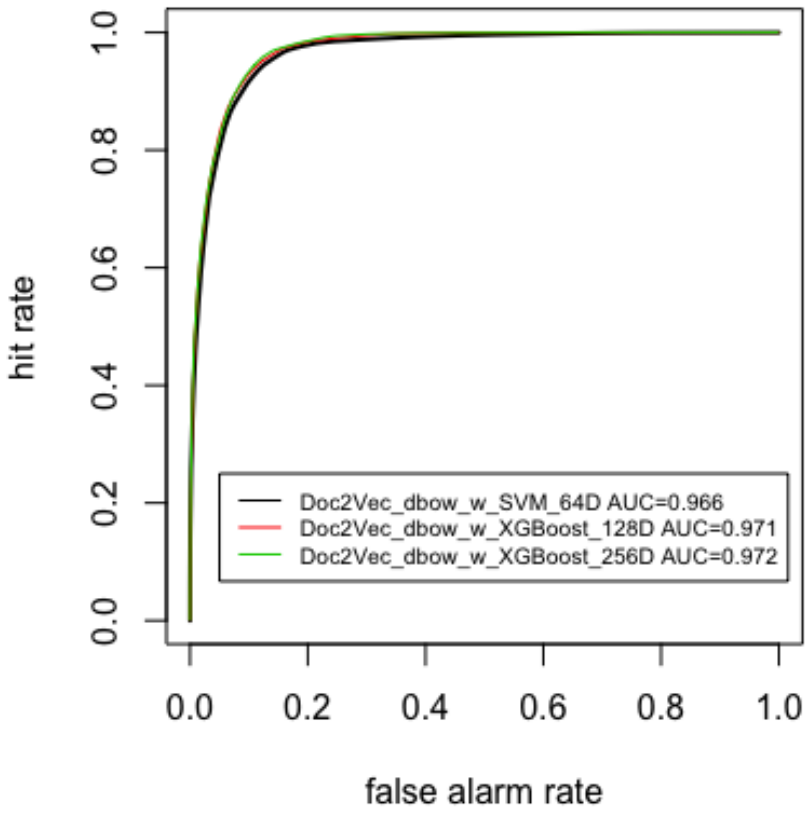
Grafike matoma tas pats rezultatas, geriausiai klasifikuojanti kombinacija išlieka ta pati, kuri buvo rasta iš ankstesnių grafikų. Geriausios kombinacijos F-matas pasikeitė nuo 0.840 iki 0.843, tai galima įvardinti kaip minimaliu pagerėjimu. Didžiausi pagerėjimai įvyko blogiausiai prognozuojančiuose vektorizavimo būduose (pvz., RP, LDA).

3.5. Geriausia detekcijos kombinacija

Iš 3.4 poskyryje pateiktų tyrimo rezultatų, buvo surastos 3 geriausios kombinacijos: 64 dimensijų atveju Doc2Vec_dbow_w vektorizavimo būdas ir tiesinis SVM modelis, 128 ir 256 dimensijų atvejais geriausios kombinacijos buvo tokios pačios, tai Doc2Vec_dbow_w vektorizavimo būdas ir gradientinio stiprinimo metodas, šiomis kombinacijomis buvo pasiektas geriausias tikslumas prognozuojant klasės žymą. Šios 3 kombinacijos toliau bus palyginamos tarpusavyje.

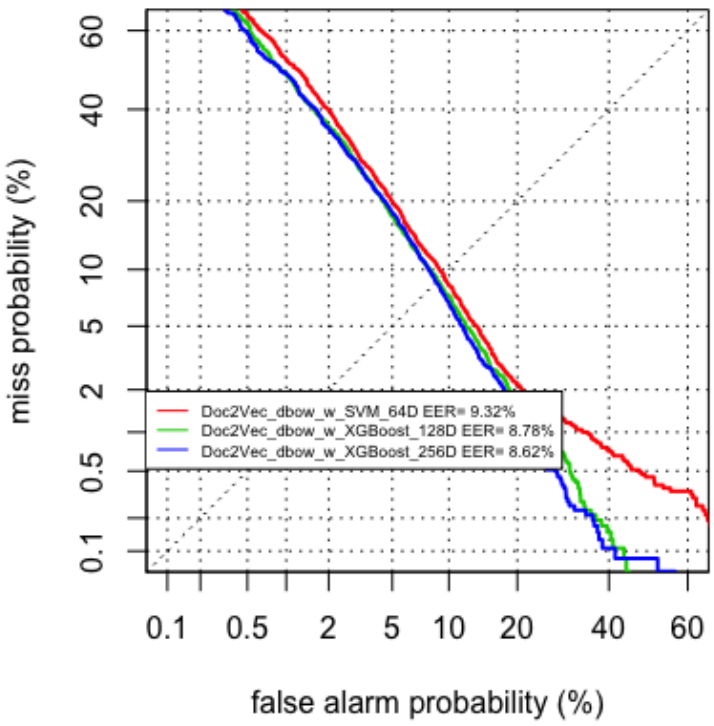
Norima detaliau norint ištirti visų geriausių 3 konstrukcijų prognozavimą, palyginimui tarpusavyje bus nubraižoma ROC, DET ir PR kreivų grafikai apskaičiuojant tikslumą įvertinančius matus bei papildomai bus pateikiama prognozuotų ir faktinių klasių proporcija – sumaišymo matrica prie EER slenksčio. Sumaišymo matricos ir prognozavimą atspindinčių rodiklių gaunamų iš šios matricos rezultatai leis identifikuoti kaip modelis prognozuoja vieną ar kitą klasę, įvertinti modelio tiksliai prognozuotų klasių procentinę dalį (angl. *accuracy*).

Apibendrinant, pateikiamos apskaičiuotos geriausių kombinacijų ROC kreivės ir AUC matai.



25 pav. Geriausių kombinacijų ROC kreivių grafikas

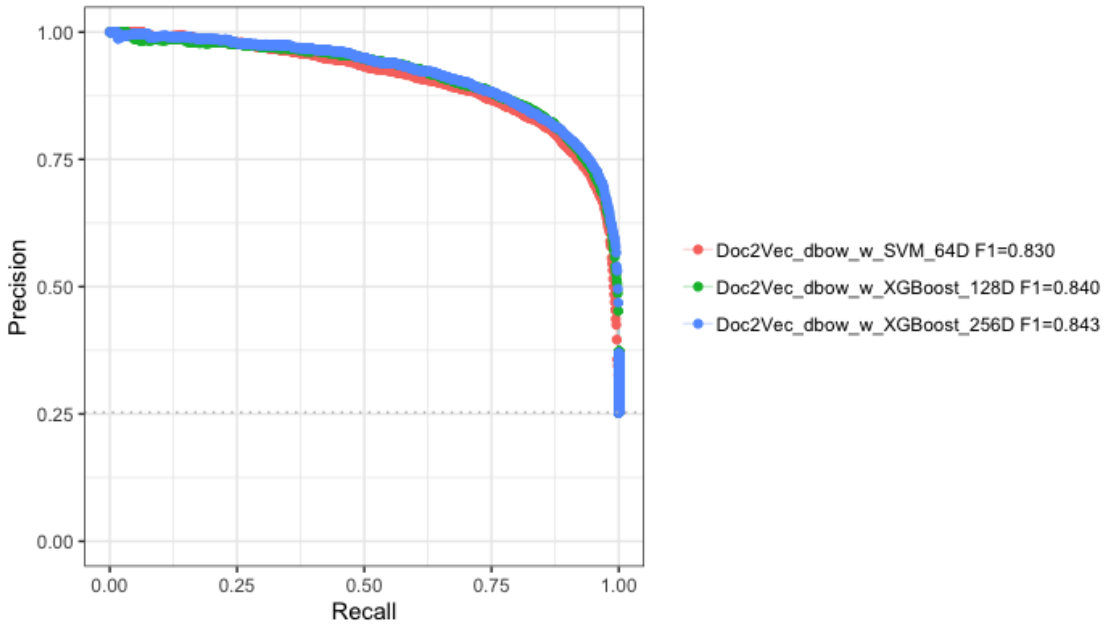
Iš ROC kreivių grafiko, matoma, kad visos kombinacijos prognozuoja panašiai tiksliai, 128D ir 256D prognozavimo tikslumo pasikeitimas minimalus ir vos pastebimas, jų AUC įverčio skirtumas tik 0.001. Didžiausias tikslumo pagerėjimo pasikeitimas buvo gautas vektorizuojant į 128 dimensijas.



26 pav. Geriausių kombinacijų DET kreivių grafikas

Iš DET kreivių grafiko įžvelgiama panaši tendencija. Pirmuoju atveju, turint 64 dimensijas suklydimo procentas buvo 9.32 proc., antruoju ir trečiuoju atveju suklydimo procentinė dalis panaši ir sudaro atitinkamai 8.78 ir 8.62 proc. Lyginant kreivių padėtis, pradžioje kreivų padėtis koordinačių erdvėje beveik identiška, tačiau pabaigoje 64D kreivėje atsiranda didesnė klaidos tikimybė (raudonos kreivės pabaigoje atsiradęs nuokrypis), kuri ir sudaro didesnę klaidingo prognozavimo procentinę dalį. Šis kreivės išsikreipimas parodo, kad yra blogiau klasifikuojama 0 klasė, t. y. teigiamą poliariškumą turintys atsiliepimai, o projekto tikslinis objektas klasė 1, prognozuojama panašiai visose kombinacijose.

Toliau pateikiamas PR grafikas, norint nustatyti tikslumo ir atkuriamumo matų priklausomybę.



27 pav. Geriausių kombinacijų PR kreivių grafikas

Grafike matyti, kad atkuriamumo ir tikslumo priklausomybė visose nagrinėjamose kombinacijose yra panaši, nepastebimas joks anomalus kitimas. Iš F-mato įverčių, identifikuojama ta pati tendencija, dimensionalumo didinimas nuo 128 iki 256 didelio tikslumo pagerėjimo nedavė.

Pateikiamos visų variantų sumaišymo matricos, kuriose bus matoma, kaip kombinacijomis sukurti klasifikatoriai prognozavo klasių žymas, kiek viso duomenų rinkinio mastu buvo teisingai ir neteisingai suklasifikuotų klasių žymų. Bus papildomai identifikuojama kurios klasės atpažinime klasifikatoriai klysta labiau – kokios rūšies klaidos būdingos prognozavimo modeliui. Pirmojo atvejo, kur atsiliepimas buvo reprezentuotas 64 dimensijų vektoriumi, geriausios kombinacijos sumaišymo matrica

10 lentelė. 64D, Doc2Vec_dbow_w ir SVM kombinacijos sumaišymo matrica

		Faktinė klasė	
		1	**0**
Prognozuojama klasė	**1**	4258	1311
	0	435	12535

Šios kombinacijos prognozavimo tikslumas sudarė 90.58 proc., matoma, kad klasifikatorius blogiau klasifikavo 0 klasės žymas nei 1. Neteisingai suklasifikuotų 0 klasės atvejai sudarė 9.47 proc. visų 0 klasės atsiliepimų, o neteisingai suklasifikuoti 1 klasės atsiliepimai sudarė 9.27 proc. visų 1 klasės atsiliepimų. Klasifikatoriaus klasių klaidos proporcijos yra panašios, tačiau projekto tikslinė klasė yra prognozuojama tiksliau. Įmonės pozicijoje, klaidingai suklasifikuoti teigiami atsiliepimai nėra tokie svarbūs, tačiau klaidingai suklasifikuoti neigiami atsiliepimai ypač svarbūs, jų neaptikimo atveju prarandama svarbi informacija norint identifikuoti neigiamas emocijas veikiančius veiksnius.

Toliau pateikiama geresnį klasifikavimo tikslumą pasiekusi antrojo atvejo su 128 dimensijomis kombinacijos sumaišymo matrica.

11 lentelė. 128D, Doc2Vec_dbow_w ir XGBoost kombinacijos sumaišymo matrica

		Faktinė klasė	
		1	**0**
Prognozuojama klasė	**1**	4275	1212
	0	418	12634

Šios kombinacijos prognozavimo tikslumas sudarė 91.21 proc., klaidingų atvejų sumažėjo klasifikuojant abi klases, t. y. papildomų dimensijų įvedimas padėjo pagerinti modelio abiejų klasių atpažinimą, o ne vienos konkrečios. Neteisingai suklasifikuoti 0 klasės atvejai sudarė 8.75 proc., o 1 klasės 8.90 proc. Ši klasifikavimo kombinacija daugiau klaidų darė prognozuodama projekto tikslinę klasę 1, tačiau bendrai klasifikavimo tikslumas padidėjo.

Toliau apžvelgiamas paskutinis atvejis, kurį sudaro 256 dimensijomis reprezentuotas atsiliepimas, pateikiama šios kombinacijos sumaišymo matrica.

12 lentelė. 256D, Doc2Vec_dbow_w ir XGBoost kombinacijos sumaišymo matrica

		Faktinė klasė	
		1	**0**
Prognozuojama klasė	**1**	4288	1195
	0	405	12651

Šios kombinacijos prognozavimo tikslumas sudarė 91.37 proc., klaidingų atvejų sumažėjo klasifikuojant abi klases, tačiau pagerėjimas nėra toks žymus kaip 128 dimensijų atveju. Neteisingai suklasifikuoti 0 klasės atvejai sudarė 8.63 proc., o 1 klasės 8.62 proc. Dimensionalumo padidinimas leido pagerinti 1 klasės prognozavimą.

Apibendrinant, iš visų pabandytų 96 skirtingų kombinacijų atvejų, buvo surasta po vieną geriausią kiekvieno dimensionalumo atveju. Visais 3 atvejais geriausiai atsiliepimą reprezentuoti ir išreikšti skaitiniu vektoriumi sugebėjo Doc2Vec_dbow_w metodas, kuris yra grįstas paskirstyto žodžių krepšelio modeliu su ‘skip-grama‘. Kaip ankščiau pateikta (žr. 2.1.2 skyrelį), metodo pranašumas, kad jis sugeba aptikti ypač retus žodžius iš didelio nestruktūrizuoto duomenų rinkinio. Projekte naudojamas duomenų rinkinys kalbiškai ir gramatiškai netvarkingas, gausus klaidų, todėl šio metodo buvimas vienu iš tiksliausių buvo potencialiai numatomas. Geriausiu klasifikavimo metodu, 2 atvejais iš 3, buvo gradientinio stiprinimo metodas (XGBoost), šis metodas yra kelių metodų sandūra, o jo sukūrimas susideda iš etapų, kai kiekviename etape bandoma prognozuoti paklaidas naujai sukuriamu modeliu. Šis modelis ir praktikoje pasižymi tikslumu, todėl taip pat buvo potencialiai numatomas būti tarp tiksliausių. Dimensionalumo didinimas patvirtina ankščiau iškeltą hipotezę apie tikslumo priklausomybę nuo dimensionalumo. Pradžioje, didėjant dimensionalumui, kombinacijų prognozavimo tikslumas gerėjo stipriau, tačiau vėliau tikslumo pagerėjimas nebuvo toks žymus. Geriausias pasiektas tikslumas yra 91.37 proc., kur EER 8.62 proc., o šios konstrukcijos dedamosios: atsiliepimo reprezentavimas 256 dimensijų vektoriumi gautu Doc2Vec_dbow_w metodu, kur klasifikavimo modelis buvo sukurtas XGBoost algoritmu.

Siekiant pagerinti duomenų rinkinio kokybę ir padidinti klasifikavimo kombinacijos tikslumą, dokumento tekstui yra atliekamas lemavimas - žodžio kamieno išskyrimas (angl. *lemmatisation*) ir pašalinamos žodžių galūnės (angl. *stemming*). Lietuvių kalba gausi skirtingų žodžių variacijų, todėl dokumente esančių žodžių pakeitimas atitinkamo žodžio bendrine forma gali pagerinti duomenų rinkinio kokybę. Norima patikrinti hipotezė, kad bendrinės formos suteikimas pagerins geriausios kombinacijos prognozavimo tikslumą. Automatiniam žodžio kamieno išskyrimui ir galūnių pašalinimui bus naudojama Lietuvių kalbos sintaksinės ir semantinės analizės informacinė sistema, kuri prieinama adresu https://semantika.lt. Šioms žodžio dalims išgauti, per šios sistemos API buvo automatiškai pateikiami žodžiai bei pasiimamos reikalingos žodinės formos. Ši procedūra užtruko apie 13 val., apie 8 proc. atsiliepimų pakeisti kalbinių dalių nepavyko, kadangi šie atsiliepimai nebuvo parašyti lietuviškomis raidėmis, todėl tokiu atveju buvo paliktas originalus tekstas. Gauti modifikuoti atsiliepimai vektorizuojami surasta tiksliausia kombinacija - Doc2Vec_dbow_w metodu, kur požymių skaičius 256. Toliau, klasifikatorius sukuriamas XGBoost algoritmu. Pateikiamas šių 2 duomenų rinkinių tikslumo palyginimas su ankščiau surastos geriausios kombinacijos tikslumu.

13 lentelė. Originalaus ir modifikuotų duomenų rinkinių tikslumo palyginimas

	Kappa	**AUC**	**EER**	**Tikslumo procentinė dalis**
Originalus d.r.	0,784	0,972	8.62 proc.	91.36 proc.
Lemuotas d.r.	0.778	0.971	8.86 proc.	91.14 proc.
Pašalintas žodžio dalies afiksas d.r.	0.773	0.968	9.09 proc.	90.91 proc.

Iš lentelės matoma, kad modifikacijų įvedimas - kalbinės bendraties rinkinio pridėjimas tikslumo nepagerino. Sukurtos naujos kombinacijos atsiliepimų poliariškumą nustatė blogiau nei surasta geriausia kombinacija su originaliu duomenų rinkiniu. Atmetama ankščiau išsikelta hipotezė dėl duomenų rinkinio kokybės pagerinimo išskiriant atsiliepimų žodžio kamieną ir pašalinant žodžio galūnę.

3.6. Geriausios kombinacijos pritaikomumas ir tolimesnės perspektyvos

Surasta geriausia kombinacija geba identifikuoti skirtingo konteksto atsiliepimų sentimento poliariškumą. Geriausia kombinacija toliau gali būti pritaikoma verslui tiriant savo klientų atsiliepimus parašytus elektroninėje erdvėje. Surasto klasifikatoriaus įdiegimas į monitoringo procesus, leistų realiu laiku, greitai identifikuoti parašyto atsiliepimo sentimento orientaciją. Lietuvos mastu dideli verslai, pvz., *pigu.lt*, turėtų galimybę ir sunaudotų mažiau resursų jeigu įsidiegtų automatinį sentimento poliariškumo atpažinimą. Tokiu būdu, klasifikatorius aptiktų neigiamo poliariškumo atsiliepimą ir iškarto informuotų atsakingus darbuotojus apie papildomų veiksmų reikalaujantį objektą. Toliau pateikiamas sukurtos kombinacijos pritaikymas naujausiuose vartotojų atsiliepimuose, kur atsiliepimai buvo gauti iš socialinio tinklo „Facebook“ bei „Circle K“ įmonės paskyros. Visi šie atsiliepimai parašyti 2019 m. gegužės mėnesį. Realus klasifikavimo kombinacijos veikimas pateikiamas lentelėje 14:

14 lentelė. Geriausia kombinacija prognozuota atsiliepimų klasė

Atsiliepimas	Prognozuojama klasė
Viskas ok, bet šiandien pirkau dešraini, prieš mane esantis zmogus pirko angliukus (vaistus)ir greit nubego i tuoleta, pardaveja ciupinejo juos, po to nenusiplovus rankų man pagamino dešraini, manau tai nehigieniska.	0.05587268
Karaliaus Mindaugo Circle K operatote Brigita,aciu uz ledukus!	0.9699397
Nemandagus personalas! Daugiau nesilankysiu	0.02041698

Matoma, kad sukurtas klasifikatorius nematytus atsiliepimus klasifikavo teisingai. Detektorius yra tinkamas tolimesniam integravimui į santykių su vartotojais vystymo sistemas. Taip pat, detektorius gali būti integruotas ne tik į CRM sistemas, bet ir į pokalbių programėles (angl. *chatbots*), kurių esmė palaikyti pokalbius, suteikti ar valdyti gautą informaciją. Tokia programėlė, kurios viena iš komponenčių būtų sukurtas detektorius, galėtų automatiškai aptikti ir valdyti klientų nusiskundimus.

Detektorius leistų išspręsti ir mažesnio masto probleminius klausimus, pvz., įmonė išleidusi naują produktą, galėtų naudojantis sukurtu detektoriumi ištirti surinktus atsiliepimus apie vartotojų patirtį, taip gaudama įžvalgų apie vartotojų nuomonės orientaciją, t. y. tokiu atveju atkreiptų dėmesį ir imtųsi papildomų veiksmų produkto gerinimui jeigu poliariškumas neigiamas. Tokio tipo vartotojų palankumo tyrimą galima pritaikyti ne tik naujai į rinką įvestoms prekėms, tačiau taip pat galima tarpusavyje palyginti keletą prekių, pvz., ištirti kuri panašaus tipo prekė yra labiau mėgstama vartotojų. Tokiai užduočiai reikėtų surinkti vartotojų atsiliepimus apie lyginamas prekes bei ištirti kurių prekių atsiliepimuose neigiama ar teigiama sentimento orientacija didesnė. Šios užduoties įgyvendinimas padėtų įmonei palyginti savo produktą su konkurentų arba išspręstų klausimą kurį produktą šalinti iš įmonės asortimento, kitaip sakant prisidėtų prie verslo vystymo ir strategijų modifikacijų. Bendrai, detektorius leidžia identifikuoti nuomonę tiriamojoje srityje, pvz., tiriama ar vartotojų neerzina siunčiamos reklaminės žinutės ir kt. Detektorius gali būti panaudojamas įvertinant ir įmonės prekės ženklo stiprumą rinkoje ar reputaciją.

Identifikuota duomenų rinkinio, kuris buvo naudojamas kuriant detektorių, probleminė sritis, tai kalbiškai netvarkingi atsiliepimai. Tolimesnės tyrimų kryptis turėtų orientuotis į duomenų rinkinio kokybės pagerinimą. Pirma, siūloma pasinaudojus neuroninių tinklų algoritmais, sumodeliuoti transformerį (*sent2sent* tipo), kuris būtų apmokytas taisyklingos lietuvių kalbos tekstais surinktais iš forumų, diskusijų portalų bei žmonių tarpusavio bendravimo erdvių. Apmokytas modelis gebėtų konvertuoti įvesties dokumentą bei išvestyje pateikti pakoreguotos rašybos dokumentą. Panašus principas naudojamas automatinių vertimų algoritmuose. Tokiam sprendimui reikalingi kompiuteriniai resursai, kurie grįsti modeliavimu ant GPU. Antra, pagerinus duomenų kokybę, naujai apmokyti geriausios kombinacijos detektorių. Siekiama, kad prognozavimo tikslumas pagerėtų 4 proc., tokiu atveju, jeigu geriausios kombinacijos tikslumas nepagerėjo, toliau orientuotis į hibridinio modelio kūrimą, įgalinant skirtingas modelių kombinacijas, pvz., XGBoost + SVM. Trečia, pasiekus norimą tikslumo pagerėjimą, detektorių naudoti neigiamo sentimento poliariškumo aptikimui ir toliau vystyti sumanesnį atsiliepimo ištyrimą. Sumanesniam neigiamo poliariškumo sentimento ištyrimui, pasinaudoti latentinį Dirichlė paskirstymą spręsti uždavinį, kuris skirtas išgauti egzistuojančias tematikas atsiliepimuose (angl. *topic modeling*), pvz., aptarnavimas ar transportavimas. Duomenų rinkinį papildyti aptiktomis tematikomis, jeigu aptikta daugiau nei viena tema, dokumentą išskaidyti. Tokiu būdu turimas klasifikavimo uždavinys, kuris orientuosis į atsiliepimo konteksto klasifikavimą. Iš verslo pusės toks klasifikatorius duos didesnę pridėtinę vertę, kadangi automatiškai atsiliepus klasifikuos pagal jų kontekstą, Įmonės turės galimybę detaliau analizuoti savo verslo problemines sritis, išskiriant neigiamą sentimento poliariškumą atsiliepimuose įtakojančius veiksnius bei pasikartojamumo dažnį.

Išvados

1. Išanalizavus literatūrą sentimentų analizės tematika, nustatyta, kad sentimento analizė yra populiarėjanti tyrimų sritis. Rezultatai gauti atlikus sentimentų analizę pritaikomi tobulinant vartotojui kuriamą patirtį skirtingose jo sąveikavimo su įmone kanaluose. Nuolatinis klientų skleidžiamos nuomonės monitoringas, įgalina sentimentų analizę atlikti realiuoju laiku ir proaktyviai reaguoti ir identifikuoti sentimento poliariškumą įtakojančius veiksnius. Sentimentų analizėje taikomi mašininio mokymosi ir leksikonu grįsti metodai, skirtingų metodų variacijas leidžia sentimentų analizę analizuoti įvairiais problematikos rakursais.

2. Sentimentų analizės tematika gana nauja tyrimų sritis Lietuvoje. Šiuo metu, tik maža dalis tyrėjų plačiau tyrinėja lietuviškus tekstus ir analizuoja vartotojų nuomonę kaip sentimentų analizės uždavinį. Konkrečių darbų susijusių su Lietuvos vartotojų nuomonės sklaida, kuriuose sentimentų analizė būtų praktiškai pritaikyta versle, nebuvo rasta.

3. Palyginus 3 skirtingus vektorizavimo dimensionalumus, 8 vektorizavimo metodus ir 4 mašininio mokymosi algoritmus, pastebėta, kad vektorizavimo metoduose dominavo paskirstyto žodžių krepšelio metodas su skip-grama, kiek mažesnis tikslumas buvo pasiektas latentinio semantinio indeksavimo ir sent2vec metodais. Nustatyta, kad dimensionalumas yra reikšmingas prognozavimo tikslumui, didelis tikslumo pagerėjimas ir progresas pastebėtas latentinio semantinio indeksavimo vektorizavimo metodo kombinacijose. Tikėtina, kad dar padidinus dimensijų erdvę, šio metodo kombinacijų tikslumas reikšmingai padidės. Skirtingas kombinacijas palyginus tarpusavyje pagal Kappa, AUC, F-matą ir EER, gradientinio stiprinimo algoritmas, dauguma atvejų, klasifikavo atsiliepimus tiksliausiai. Žodynu grįstų metodų tikslumas beveik visais atvejais buvo mažesnis už klasifikavimo tikslumą, pasiektą detekcijos modeliais. Atsitiktinių projekcijų vektorizavimo metodas atsiliepimus vektorizavimo blogiausiai ir visais atvejais prognozavimas nusileido leksikonu grįstų metodų tikslumui.

4. Surastos geriausios kombinacijos kiekvieno dimensionalumo atveju, buvo vektorizuotos vienodais metodais. Paskirstyto žodžio krepšelio metodas su skip-grama ir SVM ar XGBoost metodų kombinacijos yra tinkamiausios lietuviškų internetinių atsiliepimų klasifikavimui. Tiriant geriausias kombinacijas ir jų rezultatą, pastebėta, kad dimensionalumo reikšmingumas tarp šių kombinacijų nebuvo pastebimas, t. y. 64D ir 256D pasiektas tikslumas panašus. Papildomai įvertinus klasių klasifikavimo suklydimo procentą, geriausiu detektoriumi pasirinkta kombinacija 256 dimensijų erdvėje, o detektoriaus procentinis tikslumas sudarė 91.36 proc. Lemavimo ir žodžio dalies afikso pašalinimas, prognozavimo rezultatui įtakos neturėjo ir tikslumo nepagerino.

5. Sukurtos kombinacijos tinkamumas buvo patikrintas detektorių pritaikius keletai naujausių vartotojų atsiliepimų, kurie buvo pateikti „Facebook“ socialiniame tinkle. Nustatyta, kad klasifikatorius poliariškumą nustatė teisingai ir kombinacija tinkama tolimesnei integracijai į santykių su vartotojais palaikymo sistemas, priimant verslo vystymo sprendimus bei modifikuojant egzistuojančias strategijas - jas labiau pritaikant rinkai. Sprendimai įgalina efektyviai paskirstyti žmogiškuosius resursus, orientuojant juos į problemos sprendimą o ne aptikimą, taip pat leidžia įvertinti esamą rinką ir vartotojų poreikius.

Literatūros sąrašas

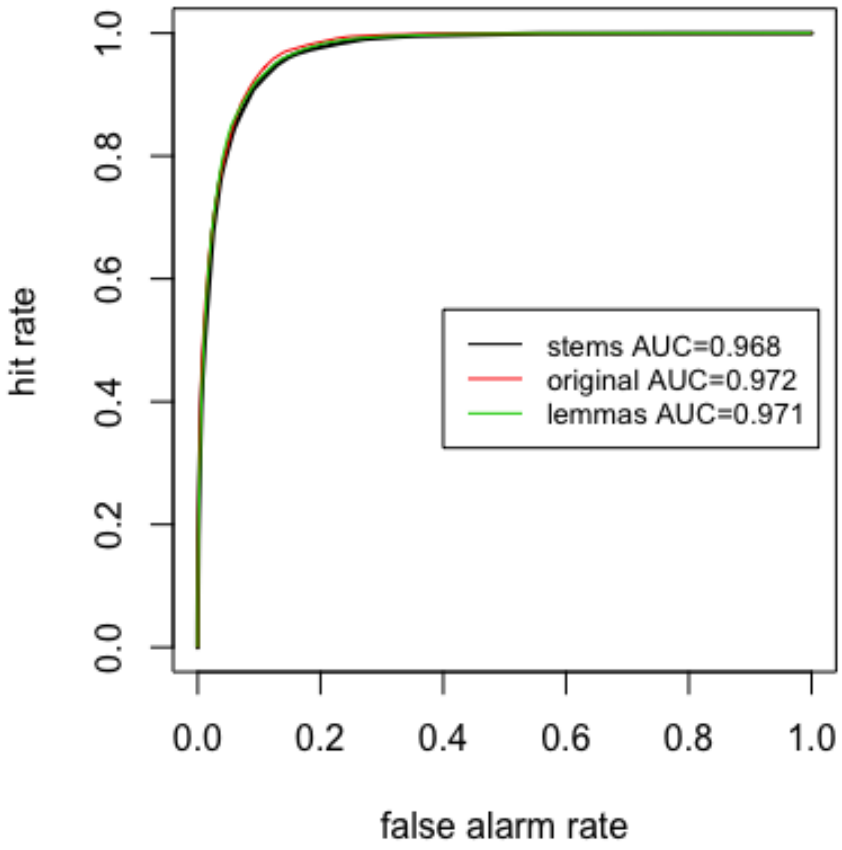
1. Maklan, S., Antonetti, P. ir Whitty, S. A Better Way to Manage Customer Experience: Lessons from the Royal Bank of Scotland. *California Management Review.* 2017. 59(2), 92-115.
2. Sharma M. ir Chaubey D.S. An Empirical Study of Customer Experience and its Relationship with Customer Satisfaction towards the Services of Banking Sector. *Journal of Marketing & Communication.* 2014. 9(3), 18-27.
3. Vavra T. G. Improving Your Measurement of Customer Satisfaction: A Guide to Creating, Conducting, Analyzing, and Reporting Customer Satisfaction Measurement Programs. Viskonsinas: ASQ Quality Press. 1997.
4. Grigoroudis, E. ir Siskos, Y. Customer Satisfaction Evaluation: Methods for Measuring and Implementing Service Quality. Londonas: Springer. 2010.
5. Roy S. Effects of customer experience across service types, customer types and time. *Journal of Services Marketing.* 2018. 32(4), 400-413.
6. Lawrence B. ir Perrigot R. Influence of Organizational Form and Customer Type on Online Customer Satisfaction Ratings. *Journal of Small Business Management*. 2015. 53, 58-74
7. Chung N. ir Kwon S. Effect of trust level on mobile banking satisfaction: A multi-group analysis of information system success instruments. *Behavior and Information Technology*. 2009. 28(6), 549-562.
8. Bustamante J. C. ir Rubio N. Measuring customer experience in physical retail environments. *Journal of Service Management*. 2017.28(5), 884-913.
9. Araffin A. A. M. ir Aziz N. A. The Effect of Physical Environment's Innovativeness on the Relationship between Hosting Quality and Satisfaction in Hotel Services. *International Journal of Trade, Economics and Finance.* 2012. 3(5), 337-342.
10. Mogaji E. ir Erkan I. Insight into consumer experience on UK train transportation services. *Travel Behaviour and Society*. 2019. 14, 21-33.
11. Gad T. Customer Experience Branding: Driving Engagement Through Surprise and Innovation. Honkongas: KoganPage. 2016.
12. Kumar A., Bezawada R., Rishika R., Janakiraman R. ir Kannan P.K. From Social to Sale: The Effects of Firm-Generated Content in Social Media on Customer Behavior. *Journal of Marketing.* 2016. 80(1), 7-25.
13. Mäntylä M. V., Graziotin D ir Kuutila M. The Evolution of Sentiment Analysis - A Review of Research Topics, Venues, and Top Cited Papers. Computer Science Review. 2018. 27, 16-32.
14. Trenz M. ir Berger B. Analyzing Online Customer Reviews - An Interdisciplinary Literature Review And Research Agenda. *ECIS 2013 Completed Research*. 2013, 83.
15. Wilson T., Wiebe J. ir Hoffmann P. Recognizing Contextual Polarity: An Exploration of Features for Phrase-Level Sentiment Analysis. *Computational Linguistics*. 2009. 5(3), 399-433.
16. Dannemann R. ir Heimann N. Social Media Mining with R. Birmingamas: Packt Publishing. 2014.
17. Liu B. (2012) Sentiment Analysis and Opinion Mining. Morgan & Claypool.
18. Cali S. ir Balaman S. Y. Improved decisions for marketing, supply and purchasing: Mining big data through an integration of sentiment analysis and intuitionistic fuzzy multi criteria assessment. *Computers & Industrial Engineering.* 2019. 129, 315-332.
19. Shayaa S., Ainin S., Jaafar N. I., Zakaria S. B., Phoong S. W., Yeong W. C., Al-Garadi, M. A., Muhammad A. ir Piprani A. Z. Linking consumer confidence index and social media sentiment analysis. *Cogent Business & Management.* 2018. 5(1).
20. Tudoran A. Why do internet consumers block ads? New evidence from consumer opinion mining and sentiment analysis. *Internet Research.* 2018. 29(1), 144-166.

21. AbdelFattah M., Galal D., Hassan N., Elzanfaly D ir Tallent G. A Sentiment Analysis Tool for Determining the Promotional Success of Fashion Images on Instagram. *International Journal of Interactive Mobile Technologies.* 2017. 11(2), 66-73.
22. Sun S., Luo C. ir Chen J. A review of natural language processing techniques for opinion mining systems. *Information Fusion.* 2016. 36, 10-25.
23. Chen Y. ir Skiena A. Building Sentiment Lexicons for All Major Languages. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics.* 2014. 2, 383-389.
24. Kaluža B. Machine Learning in Java. Birmingamas: Packt Publishing. 2016.
25. Parlar T. ir Özel S. A. An Investigation of Term Weighting and Feature Selection Methods for Sentiment Analysis. Majlesi Journal of Electrical Engineering. 2018. 12(2), 63-68.
26. Martín-Valdivia M. T., Martínez-Cámara E., Ortega J. M. P. ir López L. A. U. Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches. Journal Expert Systems with Applications: *An International Journal*. 2014. 40(10), 3934-3942.
27. Mandal S., Mahata S. K. ir Das D. Preparing Bengali-English Code-Mixed Corpus for Sentiment Analysis of Indian Languages. 2018.
28. Calderón C. A., Mohedano F. O., Álvarez M. ir Mariño M. V. Distributed Supervised Sentiment Analysis of Tweets: Integrating Machine Learning and Streaming Analytics for Big Data Challenges in Communication and Audience Research. EMPIRIA: Revista de Metodología de Ciencias Sociales. 2019. 42, 113-136.
29. Kumar A. ir Jaiswal A. Systematic literature review of sentiment analysis on Twitter using soft computing techniques. *Concurrency and Computation Practice and Experience*. 2019. 31(1).
30. Tang J., Hao S. ir Qu W. Sentiment analysis of online Chinese comments based on statistical learning combining with pattern matching. *Concurrency and Computation Practice and Experience*. Wiley. 2018. doi:10.1002/cpe.4765
31. Liu X., Wu Q. ir Pan W. Sentiment classification of micro-blog comments based on Randomforest algorithm. *Concurrency and Computation Practice and Experience*. Wiley. 2018. doi:10.1002/cpe.4746
32. Balazs J. A. ir Velásquez J. D. Opinion Mining and Information Fusion: A survey. *Information Fusion.* 2016. 27, 95–110.
33. Araque O., Zhu G. ir Iglesias C. A. A semantic similarity-based perspective of affect lexicons for sentiment analysis. *Knowledge-Based Systems*. 2019.165, 346-359.
34. Zainuddin N., Salemat A. ir Ibrahim R. Hybrid sentiment classification on twitter aspect-based sentiment analysis. *Applied Intelligence.* 2018. 48(5), 1218–1232
35. Kapočiūtė-Dzikienė J., Krupavičius A. ir Krilavičius T. A Comparison of Approaches for Sentiment Classification on Lithuanian Internet Comments. *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*. 2013. 2-11.
36. Petkevič V. Media Sentiment Analysis for Measuring Perceived Trust in Government. *Socialinis ugdymas / Socialinės komunikacijos ir pasitikėjimo sąveika*. 2018. 50(3), 23-45.
37. Kapočiūtė-Dzikienė J., Damaševičius R. ir Wozniak M. Sentiment Analysis of Lithuanian Texts Using Deep Learning Methods. 2018.
38. Dorsch I., Nikolic J., Scheibe K., Zimmer F. ir Stock W. Country-specific Sentiment on Microblogs. *Open Journal of Social Sciences.* 2018. 6, 142-158.
39. Krilavičius T., Medelis Ž., Kapočiūtė-Dzikienė J. ir Žalandauskas T. News Media Analysis Using Focused Crawl and Natural Language Processing: Case of Lithuanian News Websites. *Communications in Computer and Information Science*. 2013. 319, 48-61.
40. Kapočiūtė-Dzikienė J. ir Damaševičius R. Intrinsic Evaluation of Lithuanian Word Embeddings Using WordNet. *Artificial Intelligence and Algorithms in Intelligent Systems*. Cham: Springer. 2018.
41. Le Q. ir Mikolov T. Distributed Representations of Sentences and Documents. 2014.

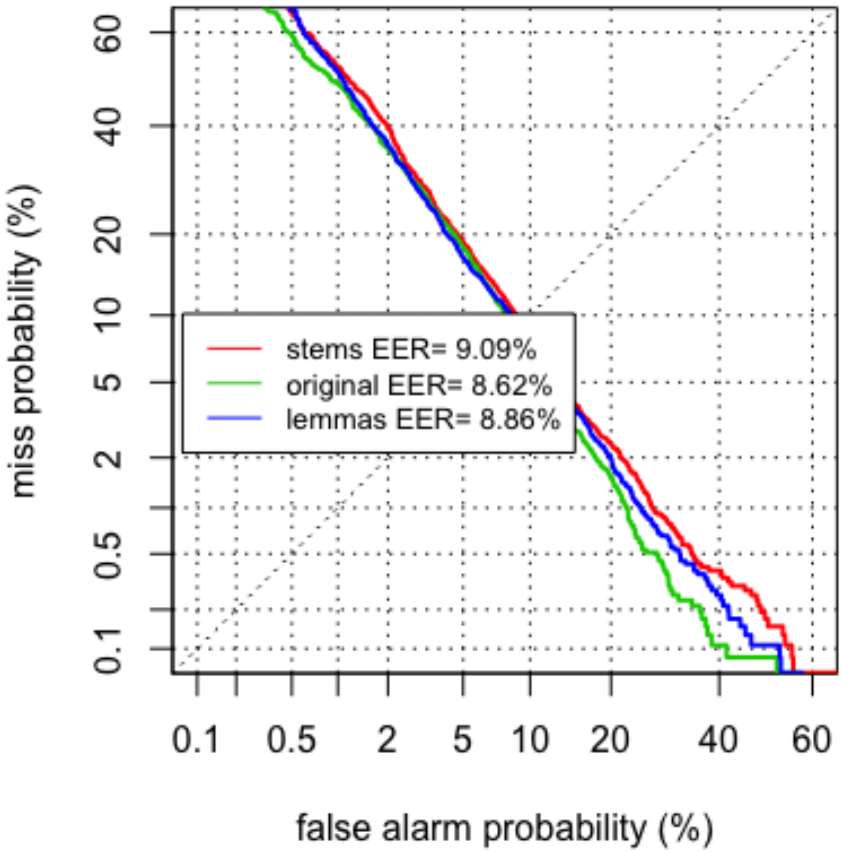
42. Ay B., Hallack I., Talo M. ir Aydin G. Evaluating deep learning models for sentiment classification. *Concurrency and Computation Practice and Experience.* 2018.
43. Kontostathis A. ir Pottenger W. M. A framework for understanding Latent Semantic Indexing (LSI) performance. *Information Processing & Management Volume*. 2006. 42(1), 56-73.
44. Yahav I., Shehory O. ir Schwartz D. Comments Mining With TF-IDF: The Inherent Bias and Its Removal. *IEEE Transactions on Knowledge and Data Engineering*. 2019. 31(3), 437-450.
45. Hoffman M D., Blei D. M. ir Bach F. Online Learning for Latent Dirichlet Allocation. *Advances in neural information processing systems.* 2010. 23, 856-864.
46. Sahlgren M. ir Karlgren J. Vector-based Semantic Analysis using Random Indexing and Morphological Analysis for Cross-Lingual Information Retrieval. *Evaluation of Cross-Language Information Retrieval Systems*. 2001. 169-176.
47. Devlin J., Chang M., Lee K. ir Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018.
48. Peng C. J., Lee K. L. ir Ingersoll G. M. An Introduction to Logistic Regression. Analysis and Reporting. *The Journal of Educational Research. 2012*, 96(1), 3-14.
49. Biau G. Analysis of a Random Forests Model. *Journal of Machine Learning Research*. 2012. 13,1063-1095.
50. Rudžianskaitė-Kvaraciejienė R. Effectiveness evaluation of public-private partnership automobile road infrastructure construction projects: daktaro disertacija. Kaunas: Technologija. 2012.
51. Srivastava D. ir Bhambhu L. Data classification using support vector machine. *Journal of Theoretical and Applied Information Technology*. 2010. 12(1), 1-7.
52. Song R., Chen S., Deng B. ir Li L. eXtreme Gradient Boosting for Identifying Individual Users Across Different Digital Devices. *Web-Age Information Management*. 2016. 43-54
53. Garcia L. P. F., Lehmann J., Carvalho A. C. P. L. F. ir Lorena A. C. New label noise injection methods for the evaluation of noise filters. *Knowledge-Based Systems*. 2019. 163, 693-704.
54. Saenz-Lechon N., Illorente J. I. G., Osma-Ruiz V. ir Gomez P. Methodological issues in the development of automatic systems for voice pathology detection. *Biomedical Signal Processing and Control*. 2006. 1(2), 120-128.

Priedai

1 priedas. Geriausios kombinacijos ir kalbinių modifikacijų ROC kreivių grafikas.



2 priedas. Geriausios kombinacijos ir kalbinių modifikacijų DET kreivių grafikas.



3 priedas. Geriausios kombinacijos ir kalbinių modifikacijų PR kreivių grafikas.