Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Tarpsritinė sentimentų analizė: mažmeninės prekybos Lietuvoje atvejis

Baigiamasis magistro studijų projektas

Anupras Kalkys

Projekto autorius

Prof. dr. Evaldas Vaičiukynas

Vadovas

Doc. dr. Aistė Dovalienė

Vadovė

Kaunas, 2025

Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Tarpsritinė sentimentų analizė: mažmeninės prekybos Lietuvoje atvejis

Baigiamasis magistro studijų projektas

Didžiųjų verslo duomenų analitika (6213AX001)

Anupras Kalkys

Projekto autorius

Prof. dr. Evaldas Vaičiukynas

Vadovas

Doc. dr. Aistė Dovalienė

Vadovė

Dr. Paulius Danėnas

Recenzentas

Prof. dr. Žaneta Gravelines

Recenzentė

Kaunas, 2025

Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Anupras Kalkys

Tarpsritinė sentimentų analizė: mažmeninės prekybos Lietuvoje atvejis

Akademinio sąžiningumo deklaracija

Patvirtinu, kad:

1. baigiamąjį projektą parengiau savarankiškai ir sąžiningai, nepažeisdama(s) kitų asmenų autoriaus ar kitų teisių, laikydamasi(s) Lietuvos Respublikos autorių teisių ir gretutinių teisių įstatymo nuostatų, Kauno technologijos universiteto (toliau – Universitetas) intelektinės nuosavybės valdymo ir perdavimo nuostatų bei Universiteto akademinės etikos kodekse nustatytų etikos reikalavimų;

2. baigiamajame projekte visi pateikti duomenys ir tyrimų rezultatai yra teisingi ir gauti teisėtai, nei viena šio projekto dalis nėra plagijuota nuo jokių spausdintinių ar elektroninių šaltinių, visos baigiamojo projekto tekste pateiktos citatos ir nuorodos yra nurodytos literatūros sąraše;

3. įstatymų nenumatytų piniginių sumų už baigiamąjį projektą ar jo dalis niekam nesu mokėjęs (-usi);

4. suprantu, kad išaiškėjus nesąžiningumo ar kitų asmenų teisių pažeidimo faktui, man bus taikomos akademinės nuobaudos pagal Universitete galiojančią tvarką ir būsiu pašalinta(s) iš Universiteto, o baigiamasis projektas gali būti pateiktas Akademinės etikos ir procedūrų kontrolieriaus tarnybai nagrinėjant galimą akademinės etikos pažeidimą.

Anupras Kalkys

Patvirtinta elektroniniu būdu

Kalkys, Anupras. Tarpsritinė sentimentų analizė: mažmeninės prekybos Lietuvoje atvejis. Magistro studijų baigiamasis projektas / vadovas prof. dr. Evaldas Vaičiukynas, vadovė doc. dr. Aistė Dovalienė; Kauno technologijos universitetas, Matematikos ir gamtos mokslų fakultetas.

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

Reikšminiai žodžiai: sentimentų analizė, tarpsritinė analizė, vektorizavimas, klasifikavimas, mašininis mokymasis, mažmeninė prekyba.

Kaunas, 2025. 65 p.

Santrauka

Šiuolaikiniame mažmeninės prekybos kontekste vis didesnę reikšmę turi vartotojų sukuriamas skaitmeninis turinys, kuriuo išreiškiamas emocinis tonas ir nusiteikimas. Siekiant išgauti įžvalgų tolimesniems strateginiams verslo sprendimams, pasitelkiama nuomonių tyryba, arba dar kitaip vadinama, sentimentų analizė. Tačiau šioje srityje analizuojant lietuvių kalbos tekstinius duomenis, susiduriama su kalbiniais ir ribotų išteklių iššūkiais. Šiame darbe lyginami klasikiniai bei modernūs vektorizavimo ir klasifikavimo modeliai, siekiant įvertinti pasirinktų metodų bei jų kombinacijų veiksmingumą, sprendžiant tarpsritinę sentimentų analizės užduotį Lietuvos mažmeninės prekybos kontekste. Darbui pasitelkiamas įvairių Lietuvos mažmeninės prekybos įmonių vartotojų tekstinių atsiliepimų rinkinys, 2011-2025 metų laikotarpiui. Atsiliepimai suskirstyti į septynias skirtingas mažmeninės prekybos sritis (*E-Marketplace*, *E-Tech*, *E- Niche*, *Groceries*, *Clothing*, *Beauty* ir *Other*).

Lyginamas klasikinis latentinės semantinės analizės (angl. *latent semantic analysis*, LSA) vektorizavimas su moderniais, transformerių architektūra grįstais vektorizavimo metodais (*Jina*, *E5*, *GTE*, *XLM-RoBERTa*), pasitelkiant reguliarizuotos logistinės regresijos, atraminių vektorių ir atsitiktinių miškų mašininio mokymosi modelius. Papildomai *XLM-RoBERTa* modelis buvo išmėgintas ir kaip atskiras klasifikatoriaus variantas sentimentų analizės užduotyje. Vektorizavimo ir klasifikavimo metodų kombinacijų sėkmingumas buvo vertintas pagal ROC AUC ir PRC AUC metrikas. Atlikus eksperimentinius tyrimus buvo išmėginti skirtingi požymių vektoriaus dimensionalumo variantai ir identifikuotas efektyviausias sentimentų analizės lietuvių kalbai sprendimas. Modernus *E5* vektorizavimas, kartu su reguliarizuotos logistinės regresijos klasifikatoriumi, apmokomas su maisto prekių srities bei testuojant su kompiuterinės technikos srities duomenimis parodė geriausius klasifikavimo rezultatus, t. y. 92,6 % tikslumas (ROC AUC = 0,978, PRC AUC = 0,998). Geriausias identifikuotas metodas, apmokytas su maisto prekių srities duomenimis, taip pat geba efektyviai klasifikuoti atsiliepimus kitose – drabužių, grožio prekių ar išskirtinėse, mažai semantiškai su apmokymo sritimi susijusiose srityse. Šie rezultatai patvirtina, kad modernūs vektorizavimo metodai gali būti efektyviai taikomi sudėtingose kalbinėse aplinkose ir įvairiose mažmeninės prekybos srityse.

Kalkys, Anupras. Cross-Domain Sentiment Analysis: A Retail Case in Lithuania. Master's Final Degree Project / supervisor prof. dr. Evaldas Vaičiukynas, supervisor assoc. prof. Aistė Dovalienė; Faculty of Mathematics and Natural Sciences, Kaunas University of Technology.

Study field and area (study field group): Applied mathematics (Mathematical Sciences).

Keywords: sentiment analysis, cross-domain analysis, vectorization, classification, machine learning, retail.

Kaunas, 2025. 65 p.

Summary

In the contemporary retail context, user-generated content is becoming increasingly more influential on consumer behavior, reflecting both emotional tone and sentiment. Aiming to extract strategic insights for further business decisions, companies rely on opinion mining, also known as sentiment analysis. Although the application of sentiment analysis is challenging in less widely spoken languages, such as Lithuanian, due to linguistic complexity and limited *NLP* resources. This thesis compares classical and modern vectorization and classification methods and their combinations, evaluating their effectiveness and applicability for cross-domain sentiment analysis within the Lithuanian retail sector. A dataset with various user text reviews for Lithuanian retail companies was used for the thesis experiments, consisting of reviews from the 2011-2025 year period. The reviews were categorized into seven different retail domains (*E-Marketplace*, *E-Tech*, *E- Niche*, *Groceries*, *Clothing*, *Beauty* and *Other*).

Comparing different vectorization methods, including the classic latent semantic analysis, *LSA*, and modern embedding methodics (*Jina*, *E5*, *GTE* and *XLM-RoBERTa*), each of these models were combined with three traditional classification methods: regularized Logistic Regression, Support Vector Machines and Random Forest. *XLM-RoBERTa* was also evaluated seperately, as a standalone classifier. These methods were assessed mainly using ROC AUC and PRC AUC metrics. After categorizing the models based on their vectorization dimensionality, the overall most effective method for Lithuanian sentiment analysis tasks was identified. The modern *E5* vectorization model, paired with the regularized logistic regression classifier achieved the best classification accuracy, e. g. 92,6 % (ROC AUC = 0,978, PRC AUC = 0,998), when trained on the grocery sector and tested on electronics store domains. The best identified method, trained on grocery domain data, also excelled, when classifying different domain reviews, such as apparel and beauty products, and even other, less semantically related testing domains. These findings confirm, that modern vectorization models can be successfully applied to low-resource and morphologically complex languages and cross-domain retail scenarios.

Turinys

Lentelių sąrašas

Paveikslų sąrašas

Santrumpų ir terminų sąrašas

BoW – žodžių krepšelio metodas;

CNN – sąsūkų dirbtiniai neuronų tinklai;

DANN – sričių adversariniai neuronų tinklai;

E2E – tiesioginio apmokymo modeliavimas.

GLU – sklendžių linijiniai vienetai;

GRU – sklendžių rekurentinių vienetų modelis;

LLM – didieji kalbos modeliai;

LoRA – žemo rango adaptavimo moduliai;

LSA – latentinė semantinė analizė;

LSTM – ilgos trumpalaikės atminties modelis;

MLM – užmaskuotas kalbos modeliavimas;

NB – Naivaus Bajeso algoritmas;

NBM – daugianario Naivaus Bajeso algoritmas;

NLP – natūralios kalbos apdorojimas;

PCA – principinių komponentų analizė;

RF – atsitiktinių miškų klasifikatorius;

RNN – rekurentiniai dirbtiniai neuronų tinklai;

RoPE – rotaciniai poziciniai įterpiniai;

SVD – singuliarių reikšmių dekompozicija;

SVM – atraminių vektorių klasifikatorius;

TF-IDF – termino dažnio – atvirkštinio dokumento dažnio metodika;

UGC – vartotojų sukuriamas turinys;

URL – universalieji adresai;

XGBoost – ekstremalaus gradiento didinimo algoritmas;

Įvadas

Šiuolaikiniame skaitmeniniame pasaulyje pastebimi ryškūs vartotojų elgesio pokyčiai, dėl vis sparčiau kuriamos subjektyvios informacijos socialinių tinklų įrašuose, internetinėse apžvalgose ir elektroninės prekybos svetainėse. Internetiniai atsiliepimai, apžvalgos bei įvertinimai tampa vieni pagrindinių šaltinių, leidžiančių vartotojams priimti sprendimą tiek internetinėje, tiek fizinėje erdvėje [8]. Lietuvos mastu verslai, analizuodami šiuos duomenis, turi galimybę suprasti savo klientų poreikius, efektyviau taikyti rinkodaros sprendimus, naujų prekių įvedimo strategijas, optimizuoti klientų aptarnavimą – tuo pačiu metu užsitikrinant ilgalaikį vartotojų lojalumą [2]. Vartotojų atsiliepimai ir įvertinimai dažnai kyla iš įvairių sričių ir platformų, todėl verslams kyla iššūkis apdorojant šiuos duomenis, bet ir užtikrinant kokybišką bei vertingą tolimesnę analizę. Įvairių sričių skirtumai, turinio konteksto variacijos bei lietuvių kalbos gramatinis sudėtingumas gali apsunkinti šį procesą. Šiame darbe analizuojami metodai, leidžiantys pagerinti vykdomos tekstinės duomenų analizės tikslumą ir pritaikomumą tarp skirtingų mažmeninės prekybos sferų.

Nuomonių tyryba, arba dar kitaip vadinama sentimentų analizė – natūralios kalbos apdorojimo (angl. Natural Language Processing, NLP) sritis, kurioje siekiama išgauti subjektyvią informaciją iš tekstinių duomenų. Šiuo būdu analizuojamas kalbos emocinis tonas, bei siekiama nustatyti vartotojo sukurtu tekstu perteikiama nuomonė [3]. Analizė gali būti grindžiama įvairiais metodais, kaip, pavyzdžiui, leksikono pagrindu veikiančiomis sistemomis, mašininio mokymosi algoritmais, giliojo mokymosi metodais ar moderniais, transformeriais grįstais didžiaisiais kalbos modeliais [12]. Sentimentų analizės efektyvumas taip pat priklauso nuo konteksto, kuriame yra sukurtas vartotojo įrašas. Tam tikri terminai ar posakiai gali turėti skirtingas reikšmes, priklausomai nuo srities ar kalbinės aplinkos [4].

Šie iššūkiai yra aktualūs ir Lietuvos mažmeninės prekybos sektoriuje, kuriame vartotojų kuriami atsiliepimai apima skirtingas produktų bei paslaugų kategorijas – nuo aprangos iki elektronikos ar maisto prekių. Kiekvienai sferai būdinga unikali kalbos specifika bei tam tikri vartotojų sentimentų požymiai. Lietuvių kalbos sudėtingumas ir riboti *NLP* ištekliai mažina sentimentų analizės modelių efektyvumą bei jų adaptavimo potencialą skirtingose srityse, ypač lyginant su anglų ar vokiečių kalbomis [1]. Tai kelia būtinybę palyginti ir įvertinti modernias metodikas, kuriomis gebama apeiti kalbinių kliūčių ribojimus ir adaptuoti naudojamus sentimentų analizės modelius prie skirtingų sričių turinio ir konteksto variacijų.

Tokiu būdu pabrėžiama tarpsritinės sentimentų analizės metodikų svarba, siekiant spręsti problemas, susijusias su ribotais įvairių sričių duomenų ištekliais ir turinio variacijomis lietuvių kalboje. Nors tarpsritinė sentimentų analizė yra plačiai nagrinėjama anglakalbėse rinkose, šiame darbe išbandomi ir vertinami pažangūs modeliai ir metodikos, siekiant juos pritaikyti Lietuvos mažmeninės prekybos sektoriuje – sferoje, kurioje egzistuoja platus sričių skirtumų ir kalbinių niuansų spektras. Šiuo darbu siekiama prisidėti ne tik prie sentimentų analizės metodikų plėtros ir pritaikymo lietuvių kalbai, bet ir atskleisti naujas verslo analitikos perspektyvas, skatinant inovatyvius ir efektyvius, duomenimis grįstus sprendimus mažmeninės prekybos sektoriuje.

Darbo problema: Lietuvos mažmeninės prekybos sektoriuje, vykdant sentimentų analizę susiduriama su skirtingų sričių kalbiniais ypatumais ir ribotais lietuvių kalbos *NLP* ištekliais. Siekiant panaudoti sentimentų analizės metodus skirtingose srityse, reikalingi efektyvūs metodai, gebantys įveikti kontekstines ir kalbines variacijas.

Darbo tikslas: Tarpsritinės sentimentų analizės metodikos lietuvių kalbai gerinimas.

Darbo uždaviniai:

1. atlikti literatūros apžvalgą ir pagrįsti sentimentų analizės Lietuvos mažmeniniame sektoriuje svarbą ir problematiką.
2. apžvelgti atliktus sentimentų analizės tyrimus Lietuvoje.
3. pasirinkti modelius ir metodus sentimentų analizės uždaviniui, papildant ir pernaudojant lietuvių kalbos atsiliepimų „evertink.lt“ ir „Facebook“ platformose surinktą duomenų rinkinį.
4. atlikti empirinį tyrimą, kuriame būtų įvertinta pasirinktų modelių veiksmingumas analizuojant sentimentus tarp skirtingų sričių.
5. įvertinti tyrimo rezultatus, pateikiant jų interpretaciją ir rekomendacijas, kaip tarpsritinės sentimentų analizės sprendimus galima pritaikyti verslo praktikoje, siekiant pagerinti vartotojų elgesio įžvalgas ir sprendimų priėmimą mažmeninės prekybos įmonėse.

1. Literatūros apžvalga

Mažmeninės prekybos įmonėms siekiant suprasti bei nuspėti vartotojų elgesį, nustatyti atsirandančias tendencijas rinkoje bei tobulinti verslo strategijas, klientų atsiliepimai yra vienas svarbiausių galimų veiksnių. Vis sparčiau augantis aktyvumas skaitmeninėse platformose skatina vartotojus kurti bei tuo pačiu metu naudoti kitų vartotojų sukuriamą turinį (angl. *User Generated Content*, UGC), pavyzdžiui, internetines apžvalgas, atsiliepimus ar socialinių tinklų įrašus – didinant šio turinio prieinamumą bei svarbą. Ženklus *UGC* skaičiaus augimas suteikia verslams didėjantį kiekį informacijos, leidžiančios gauti įžvalgų apie vartotojų pageidavimus, norus, nepasitenkinimus bei atsirandančias tendencijas. Teigiama, jog klientų atsiliepimai skaitmeniniame kontekste tapo vienas esminių veiksnių vartotojams ir verslams priimant sprendimus – akcentuojant perteikiamo emocinio tono suvokimo svarbą tarp vartotojų sukurtų tekstinių duomenų [3].

Mažmeninės prekybos įmonių sėkmė priklauso nuo jų gebėjimo interpretuoti bei adaptuotis prie kintančių klientų poreikių. Sentimentų analizės pagalba, verslams suteikiama galimybė imtis unikalios iniciatyvos, siekiant užsitikrinti pranašumą konkurencingose rinkose. Įvairios gaunamos įžvalgos, leidžiančios suvokti klientų nusiteikimą bei emocijas, įgalina verslus pagrįstai adaptuoti savo kuriamus produktus, paslaugas ar strategijas, tokiu būdu gerinant pelningumą, vartotojų patirtį bei ilgalaikį lojalumą [4].

Skaitmeniniai klientų atsiliepimai gali teikti ne vien paviršutinišką ir tiesioginį grįžtamąjį ryšį, taipogi, pasinaudojant *UGC*, verslams suteikiama galimybė daryti ilgalaikę įtaką vartotojų pasirinkimams bei formuoti pozityvų prekės ženklo vardą. Lietuvos mastu, pažangios didžiųjų kalbos modelių sentimentų analizės metodikos, pavyzdžiui *BERT* (angl. *Bidirectional Encoder Representations from Transformers*), ar jos modifikacijos, prisitaiko prie sudėtingų lietuvių kalbos lingvistinių niuansų. Siekiant geriau suprasti savo klientus, numatyti paklausos pokyčius, bei optimizuoti savo siūlomas prekes ar paslaugas, efektyvus *UGC* panaudojimas tampa vienu svarbiausiu uždavinių šiandieninėse mažmeninės prekybos rinkose, ką gali suteikti sentimentų analizė [11].

1.1. Sentimentų samprata ir jų analizės svarba

Sentimentas, kaip terminas, simbolizuoja asmens perteikiamą požiūrį, vertinimą ar emocinę būseną ties tam tikrais objektais ar idėjomis. Lingvistikoje, ši sąvoka aprėpia teigiamus, neigiamus ar neutralius tonus, atsispindinčius žodžiuose, taip atvaizduojant kalbėtojo asmeninę poziciją ar jausmus [3]. Tai yra esminis žmogiško bendravimo konstruktas, neatsiejamas komunikuojant, ar priimant kasdienius sprendimus. Šiuolaikinei visuomenei vis labiau natūraliai orientuojantis į skaitmenizaciją įvairiose gyvenimo srityse, sentimentai, užfiksuoti tekstiniu formatu, tapo vienu svarbiausių šaltinių verslams, siekiantiems suprasti vartotojų elgesį bei atitinkamai derinti verslo strategijas. Literatūroje teigiama, jog sentimentų analizė yra glaudžiai susijusi su pasaulyje vykstančiomis tendencijomis, kurias tyrinėja elgsenos ekonomikos (angl. *behavioral economics*) mokslas. Ši sritis nagrinėja, kaip emocijos, psichologinės nuostatos bei kognityvinis šališkumas (angl. *cognitive bias*) daro įtaką vartotojų sprendimams ir ekonominei veiklai. Vartotojų sentimentai, išreiškiami per internetinius atsiliepimus ar socialinių tinklų įrašus padeda suprasti ne visada racionalius, tačiau dažnai, stipriai emocijomis pagrįstus sprendimus. Pavyzdžiui, vartotojo teigiamas emocinis atsakas į tam tikrą prekės ženklą, ar prekės ženklo perteikiamas vertybes gali skatinti vartotojo lojalumą, netgi jei alternatyvos

yra ekonomiškai naudingesnės. Tokiu būdu sentimentų analizė gali būti panaudojama ne tik produktų ar paslaugų įvertinimui, tačiau ir gilesniam vartotojų elgesio modeliavimui [5].

Versle, sentimentai traktuojami kaip esminė priemonė, kurios pagalba interpretuojami vartotojų atsiliepimai bei rinkos dinamika. *UGC*, kurį gali sudaryti internetiniai atsiliepimai, įvertinimai ar socialinių tinklų įrašai, tapo vienu pagrindiniu šaltiniu, vykdant sentimentų analizę. Tinkamai panaudojant šiuos duomenis, verslams atsiranda galimybė geriau prisitaikyti prie rinkos pokyčių. Kaip pavyzdys, pastebimi teigiami sentimentai apie verslo vykdomą tvarią veiklą gali nurodyti potencialą didinti įmonės siūlomą ekologišką produkciją, o neigiami atsiliepimai apie klientų aptarnavimą gali padėti nustatyti problematiškas sritis vykdomoje veikloje [1]. Toliau pateikiamas termino „sentiment analysis“ populiarumas internetinės paieškos sistemoje „Google“ (žr. 1 pav.)

1 pav. Termino „sentiment analysis“ populiarumo pokytis „Google“ paieškoje nuo 2004 metų [Šaltinis: „Google Trends“]

Vienas iš pagrindinių sentimentų analizės privalumų slypi galimybėje atskleisti vartotojų suvokimą bei nusiteikimą išlaidauti. Teigiami sentimentai, skirti specifiniam produktui ar paslaugai, dažnu atveju siejasi su aukštesne suvokiama verte, tokiu būdu padidinant vartotojo išlaidavimo potencialą. Pavyzdžiui, vartotojai gali sieti ekologiškai tvarią produkto pakuotę, aukštos kokybės medžiagas ar prekės ženklo prestižą su didesne verte, kas yra matoma išreiškiamuose sentimentuose. Tai pastebima didėjančiuose vartotojų noruose išlaidauti produktams, kurie pristatomi kaip tvarūs ar atitinkantys modernius etinius standartus. Atliktame Aschemann-Witzel ir Zielkės [6] tyrime teigiama, jog vartotojai perteikia intensyvesnius teigiamus sentimentus bei didesnį norą išlaidauti, kai produktai turi „*organiškas*“ (angl. *organic*) arba „*sąžiningas*“ (angl. *fair trade*) žymėjimus. Šis reiškinys pastebimas ir tada, kai būtent taip pažymėtų bei paprastesnių, nežymėtų alternatyvių produktų skirtumas yra minimalus. Panašiai, Merbah ir Benito-Hernández [7] atliktame tyrime taip pat teigiama, kad vartotojai Ispanijoje yra pasirengę mokėti daugiau už kavą, pažymėtą „*sąžiningas*“ ar „*UTZ*“ žymėjimais, pabrėžiant, kad tokie ženklinimai didina vartotojo suvokiamą produkto vertę bei skatina vartotoją pirkti. Šie du pavyzdžiai parodo, jog vartotojo produkto vertės suvokimas dažnai gali priklausyti nuo produkto inovatyvumo bei tvarumo aspektų perteikimo. Kita vertus, vartotojų nepasitenkinimas produkto kokybe, aptarnavimu ar kiti neigiami sentimentai gali sumenkinti suvokiamą vertę, tiesiogiai darant įtaką vartotojo išlaidavimo sprendimams. Tinkamas sentimentų analizės panaudojimas leidžia išgauti įžvalgas, kurios gali suteikti pranašumą konkurencingose rinkose. Identifikuojant esmines produktų savybes ar funkcijas, teikiančias vartotojams pasitenkinimą, bei kartu pasitelkiant kainodaros strategijų korekcijas, sentimentų analizės pagalba gali būti efektyviai optimizuojami verslo planai [8].

Sentimentų analizė yra ne tik priemonė tiesioginių, momentinių vartotojų problemų sprendimui, tačiau ir strateginis įrankis ilgalaikiam verslo planavimui. Analizuojant sentimentų tendencijas

skirtinguose klientų segmentuose ir produktų kategorijose, atsiranda galimybė numatyti vartotojų elgesio pokyčius bei numatyti rinkos pokyčius. Kapočiūtės-Dzikienės, Damaševičiaus ir Wožniak'o [1] teigimu, Lietuvos mažmeninės prekybos sektoriuje, kultūriniams bei lingvistiniams niuansams darant didžiulę įtaką vartotojų elgesiui, sentimentų analizė tampa vienu svarbiausiu įrankiu, leidžiančiu verslams atitikti savo klientų lūkesčius, teisingai pritaikant vykdomo verslo strategijas. Be to, sentimentų analizė leidžia nuodugniai valdyti prekės ženklą. Analizuojant sentimentų pokyčius laike, verslai gali užkirsti kelią potencialioms reputacinėms krizėms. Stebint sentimentų pokyčius realiu laiku, įmonės gali atitinkamai vertinti reikšmingų įvykių, kaip naujo produkto išleidimo ar įvairių skandalų poveikį verslui, taip suteikiant galimybę kontroliuoti ir išlaikyti vartotojų pasitikėjimą [9].

Siekiant praktinio pritaikymo, sentimentų analizei įvykdyti reikalingas vartotojų išreiškiamų emocijų poliariškumo nustatymas, leidžiantis klasifikuoti vartotojų tekstines išraiškas. Verslai, įvertindami, ar analizuojamas sentimentas yra teigiamas, neigiamas ar neutralus, gali išgauti tolimesnes išvadas iš šių nestruktūrizuotų duomenų, bei geriau suprasti vartotojų elgesį skatinančias emocijas.

1.1.1. Sentimentų poliariškumas

Sentimentų poliariškumas įvardijamas kaip išreikštos vartotojo nuomonės tono klasifikavimas į teigiamą, neigiamą ar neutralią kategorijas. Toks sudėtingų emocinių išraiškų supaprastinimas yra vienas esminių sentimentų analizės bruožų, leidžiančių išgauti vertingų įžvalgų iš tekstinių duomenų [3]. Sentimentų poliariškumo konceptas siejamas su semantinės orientacijos teorija, pagal kurią tam tikri žodžiai ir frazės yra labiau siejami su konkrečiais emociniais tonais [10]. Pavyzdžiui, atsiliepimas internetinėje parduotuvėje: „*Viskas puiku, tik labai ilgas laiko tarpas tarp užsakymo ir įvykdymo*“ parodo teigiamą sentimentą produktui, tačiau neigiamą sentimentą nukreiptą į užsakymo vykdymą. Tokiu būdu poliariškumo detekcija įgalina verslus identifikuoti savo vykdomos veiklos privalumus ir trūkumus.

Vertinant vartotojų pasitenkinimą, aiškinantis problematiškas verslo sritis ar adaptuojant verslo strategijas, poliariškumo analizė yra kertinis elementas. Supaprastintas sentimentų kategorizavimas leidžia sudaryti pamatus tolimesnei, kiekybinei analizei. Siekiant suprasti vartotojų norus, pageidavimus bei prognozuojant tolesnę rinkos tėkmę, sentimentų analizė gali atskleisti produktus, sukeliančius teigiamas vartotojų reakcijas, įgalinant įmones suteikti pirmenybę produktams ar paslaugoms su didžiausios grąžos potencialu. Globaliu mastu, poliariškumo detekcija verslams leidžia vertinti vartotojų sentimentus, siekiant prisitaikyti prie kultūrinių ar ekonominių skirtumų per įvairias rinkas. Sentimentų variacijos, priklausomos nuo pasaulio vietovių, gali reikalauti skirtingų rinkodaros ar kainodaros metodų, tad tokiu būdu užtikrinamas stabilus konkurencingumas pasaulinės ekonomikos atžvilgiu [10] [11].

Tačiau reikalinga paminėti kylančius iššūkius – lingvistinės subtilybės, kaip idiomai (t. y., tik kuriai nors kalbai būdingas vientisos reikšmės sustabarėjęs žodžių junginys, kurio reikšmė nesutampa su jį sudarančių žodžių reikšmėmis, turintis savitą leksinę reikšmę), sarkazmas ar kultūriniai niuansai, dažnai užgožia tikrąjį teiginio sentimentą. Pavyzdžiui, kitas atsiliepimas internetinėje parduotuvėje: „*Tiesiog nuostabios kojinės – skylė atsirado vos po vieno skalbimo*“ perteikia neigiamą sentimentą, nepaisant tariamai teigiamos pirminės frazės dalies. Tokie niuansai gali kelti kliūčių analizėje užfiksuojant latentines (paslėptas) vartotojų emocijas, kurios tekste nėra akivaizdžios [9] [11].

Nors poliariškumo detekcija supaprastina sentimentų analizę iki aiškių, klasifikuotų nuomonių, reikalingas tolimesnis detalizavimas, siekiant efektyvesnių įžvalgų – analizuojant sentimentus skirtingais lygmenimis. Galimybė analizuoti sentimentus bendru, dokumento lygmeniu, skirstant sakiniais, frazėmis ar specifiniais požymiais, suteikia dar geresnę galimybę įsigilinti į vartotojų išreiškiamas emocijas [9].

1.1.2. Sentimentų analizės lygmenys

Sentimentų analizė gali būti atliekama keturiais pagrindiniais lygmenimis, priklausomai nuo reikalingų įžvalgų ar turimų išteklių. Dokumento, sakinio, frazės bei aspekto lygmenys gali suteikti unikalių įžvalgų verslams, nuo plataus masto rinkos tendencijų, iki itin smulkių detalių specifinėse produkto savybėse, siekiant išsamesnės informacijos tolimesnei, sėkmingai verslo veiklai [12]:

1. **Dokumento lygmens analizė**. Šis metodas traktuoja tekstą kaip vieną, singuliarų objektą, siekiant identifikuoti apibendrintą vartotojo sentimentą. Pavyzdžiui, internetinis atsiliepimas: „*Labai patenkintas, informatyviai ir labai greitai*“, būtų klasifikuojamas kaip teigiamas sentimentas. Šio lygmens analizė naudinga, siekiant apibendrinti bei klasifikuoti bendrą vartotojo sukurtą tekstinę išraišką. Tokiu būdu daroma prielaida, jog vartotojo sentimentas per visą tekstą išlieka vienodas bei nesikeičiantis – tačiau neretai, išreiškiamos emocijos tame pačiame tekste kinta, susiduriant su potencialiais netikslumais galutiniame vertinime [13];
2. **Sakinio lygmens analizė**. Tekstas suskaidomas į individualius sakinius, norint išgauti išsamesnių įžvalgų. Pavyzdžiui, atsiliepimas: „*Viskas neįtikėtinai greitai. Vienintelis trūkumas, kad VENIPAK sakė dirbantys tik iki 5 val*.“, skaidomas į du sakinius. Vienas iš jų, kaip teigiamas sentimentas „*viskas neįtikėtinai greitai*“, išskiriant „*neįtikėtinai greitai*“ kaip pagrindinį indikatorių. Kitas sakinys, kaip neigiamas sentimentas, pateikiamas „*Vienintelis trūkumas, kad VENIPAK sakė dirbantys tik iki 5 val.*“, išskiriant „*vienintelis trūkumas*“ bei „*dirbantys tik iki 5 val.*“ kaip esminius indikatorius. Šio lygmens analizė naudinga, siekiant išskirti konkrečias vartotojų pasitenkinimo bei nepasitenkinimo sritis [9];
3. **Frazės lygmens analizė**. Analizuojant tekstą šiuo metodu, dėmesys sutelkiamas į individualias sakinio dalis ar frazes, kurios paaiškina tam tikras subtilybes, praleidžiamas su platesne sakinio lygmens analize. Atsiliepime „*Maistas buvo puikus, tačiau aptarnavimas tragiškas*“, pastebimos dvi frazės: „*Maistas buvo puikus*“, identifikuojamas kaip teigiamas sentimentas, kur pagrindinis indikatorius yra „*puikus*“. Antroje frazėje, „*aptarnavimas tragiškas*“, išskiriamas „*tragiškas*“ kaip neigiamo sentimento indikatorius. Šio lygmens analizė gali suteikti detalesnių įžvalgų nei sakinio lygmuo, dėl efektyvios sentimentų atskirties net tada, kai viename sakinyje išreiškiami skirtingi požiūriai. Pritaikant šį lygmenį atsiliepimų ar socialinių tinklų srityse, kuriuose gausu trumpų, fragmentuotų posakių, atsiranda galimybė išgauti informatyvesnes įžvalgas [14];
4. **Aspekto lygmens analizė**. Tekstas nagrinėjamas detaliausiu lygiu, daugiausia dėmesio skiriant specifiniams produkto ar paslaugos požymiams. Pavyzdžiui, atsiliepime: „*Laba diena, Prekės pristatytos greitai. Prekės kokybė nebloga, išskyrus tai, kad pietų servize viena maža lėkštutė po puodeliu buvo sudužusi*“, išskiriami trys atskiri požymiai. Išskiriant pristatymo greitį kaip požymį, susiejant su teigiamu sentimentu – „*Prekės pristatytos greitai*“; išskiriant bendrą produkto kokybę, taip pat susiejant su teigiamu sentimentu – „*Prekės kokybė nebloga*“, išskiriant „*greitai*“ ir „*nebloga*“, kaip pagrindinius indikatorius. Kitas aspektas, specifinė produkto problema, siejama su neigiamu sentimentu – „*išskyrus tai, kad pietų servize viena maža lėkštutė po puodeliu buvo sudužusi*“, paskiriant „*sudužusi*“ kaip pagrindinį indikatorių. Aspekto lygmens sentimentų analizė

yra vertinga konkurencingose srityse, kur ypač svarbu įsigilinti į tam tikrų vartotojų nuomones bei pageidavimų specifikas [4].

Kiekvienas sentimentų analizės lygmuo sprendžia skirtingus verslo uždavinius. Dokumento lygmens analizė efektyviai suteikia plataus masto sentimentų apibendrinimą, sakinio bei frazės lygmens analizės išskiria tarpines, tačiau vertingas žinias, o aspekto lygmens analizė išryškina smulkias, individualias produkto arba paslaugos savybes. Tačiau kiekvienas sentimentų analizės lygmuo kelia savotiškų iššūkių. Dokumento lygmens analizė gali atsitiktinai praleisti esmines vartotojo nepasitenkinimo sritis. Sakinio lygmens metodai gali teikti netikslius rezultatus, tyrinėjant sudėtingus bei poliariškai mišrius sakinius [9]. Frazės lygmens metodikos yra sunkiau pritaikomos kalbose su komplikuota gramatika, kuriose nėra griežtos struktūros. Aspekto lygmens analizei reikalingi sudėtingesni įrankiai bei metodai, siekiant tiksliai apdoroti tam tikrų, konkrečių savybių atsiliepimus [14] [4].

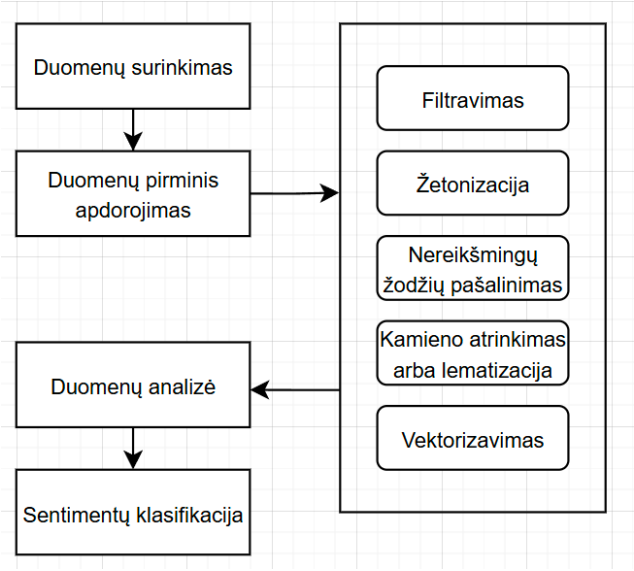
Sentimentų analizei pritaikomi ir hibridiniai metodai, jungiantys kelis lygmenis, siekiant geresnių įžvalgų. Praktikoje, dokumento lygmens analizė gali pateikti apibendrintą sentimentų kryptį visame tekste, o aspekto lygmens analizė identifikuoja specifines savybes, skatinančias būtent tokį sentimentų analizės kryptingumą. Hibridiniai metodai yra naudingi mažmeninės prekybos kontekste, kur vartotojų sentimentai gali būti mišrūs tame pačiame tekste. Panaudojant minėtą hibridinį metodą, pavyzdžiui, sujungiant dokumento ir aspekto lygmens analizę, įmonėms atsiranda galimybė vertinti tam tikro produkto efektyvumą rinkoje, ir kartu sužinoti smulkias, tačiau itin svarbias vartotojų išreiškiamas detales [15].

Apibendrinant, sentimentai, kaip samprata, vaizduoja asmens subjektyvią nuomonę, nusiteikimą ir emocijas, analizėje kategorizuojamas kaip teigiamas, neigiamas ar neutralias. Sentimentai naudingi analizuojant *UGC*, gerinant verslo strategijas, prognozuojant rinkos pokyčius bei valdant prekės ženklo reputaciją. Sentimentų poliariškumas supaprastina sudėtingas vartotojų perteikiamas emocines išraiškas į suprantamas kategorijas, o išsamesniam vertinimui, sentimentai gali būti analizuojami dokumento, sakinio, frazės arba aspekto lygmenimis, priklausomai nuo konkretaus tikslo. Tolesniame skyriuje apžvelgiamos pagrindinės sentimentų analizės metodologijos, kuriomis naudojantis vykdomas sentimentų analizės klasifikavimas bei vertinimas.

1.2. Sentimentų analizės metodologijos

Sentimentų analizė – viena sparčiausiai besivystančių *NLP* sričių, verslams vis labiau atsižvelgiant į *UGC* bei jų suteikiamas įžvalgas. Dėl to atsiranda didelė paklausa efektyviems sentimentų analizės metodams. Šiame skyriuje nagrinėjamas teksto pradinis apdorojimas bei esminės sentimentų analizės metodikos.

Sentimentų analizės procesas prasideda nuo pirminio duomenų apdorojimo, dėl to, kad naudojami tekstiniai duomenys dažniausiai savyje turi triukšmo, kas sukelia kliūčių tolimesniam, efektyviam modeliavimui. Pagal analizuojamą literatūrą išskiriamos metodikos, kurias naudojant siekiama pagerinti analizės rezultatus, t. y., filtravimas, žetonizacija (angl. *tokenization*) (teksto skaidymas į smulkesnius vienetus), nereikšmingų žodžių (angl. *stop words)* pašalinimas, kamieno atrinkimas (angl. *stemming*) ar lematizacija (angl. *lemmatization*) bei vektorizacija [16] [17]. Žemiau pateikiama supaprastinta sentimentų analizės eiga (žr. 2 pav.).



2 pav. Pradinė, klasikinė sentimentų analizės eiga (adaptuota į lietuvių kalbą pagal [18])

Pirminis procesas – duomenų triukšmo pašalinimas, arba kitaip, filtravimas, yra vienas svarbiausių sentimentų analizės etapų, kuris užtikrina vėlesnį modelio efektyvumą ir tikslumą. Sistemiškai pašalinami arba transformuojami nereikalingi, klaidinantys arba dar kitaip, triukšmingi elementai tekste. Pavyzdžiui, universalieji adresai (angl. *Uniform Resource Locators*, URL), hiperteksto ženklinimo kalbos (angl. *HyperText Markup Language*) žymės, ar kiti vartotojų paliekami simboliai, pavyzdžiui, jaustukai (angl. *emoticon*). Tokiu būdu užtikrinama pastovi bei nuosekli struktūra duomenyse [16].

Žetonizacijos procesas – teksto skaidymas į mažesnius vienetus (angl. *tokens*), kuris leidžia sentimentų analizės modeliams geriau apdoroti bei analizuoti tekstą. Literatūroje kaip efektyvesnis metodas, minima subžodžių žetonizacija (angl. *subword tokenization*) – šis metodas leidžia efektyviau apdoroti sudėtingesnes kalbas ir retus žodžius. Minimi baitų porų kodavimo (angl. *byte pair encoding*), *WordPiece* bei *SentencePiece* metodai užtikrina semantinių niuansų išlaikymą [19] [20].

Nors subvienetų žetonizacija pritaikoma daugumai kalbų, tačiau ypatingą reikšmę šis metodas turi pritaikant morfologiškai sudėtingoms kalboms, tokioms kaip suomių ar lietuvių. Pavyzdžiui, šiais metodais, žodis „*nepaprastai*“ gali būti žetonizuojamas į tris dalis:

1. „*ne*“, perteikiantis neigimą;
2. „*paprast*“, išskiriama žodžio šaknis, simbolizuojanti „*nesudėtingas*“ arba „*niekuo neišsiskiriantis*“;
3. „*-ai*“, priesaga, pakeičianti žodį į prieveiksmį.

Žetonizacijos būdu modelis geba efektyviau generalizuoti įvairiuose kontekstuose, kaip nuo žodžio „*paprastas*“ į „*nepaprastas*“, ir iš „*paprastai*“ į „*nepaprastai*“, atitinkamai užfiksuojant kalbinius šablonus bei išlaikant semantinius ir sintaksės ryšius [20].

Kitame žingsnyje pašalinami nereikšmingi žodžiai, t. y., dažnai pasikartojantys, pertekliniai bei minimalią semantinę vertę tam tikroje kalboje turintys žodžiai. Tokiu būdu sumažinus bendrą žodžių kiekį analizuojame tekste, pagerinamas modelio našumas, leidžiant modeliui atlikti skaičiavimus tik ant esminių, kokybiškų bei įtaką galutinei analizei galinčių daryti elementų [16].

Pašalinus nereikšmingus žodžius, atliekamas kamieno atrinkimas arba lematizacija – tekstinių duomenų normalizavimo metodikos. Abejais metodais siekiama sumažinti žodžius iki paprastesnės formos, atveriant kelią tolimesniems sentimentų analizės procesams. Nors abu metodai naudojami tuo pačiu tikslu, kamieno atrinkimas – primityvesnis žodžio sumažinimas, t. y., taisyklėmis grįsta euristinė metodika, atskeliant žodžio galūnę. Taip atliekant normalizavimą, kartais gaunami nestandartiniai žodžio sumažinimai, t. y., žodis „*mylėjau*“ gali būti sutrumpinamas iki „*mylėj*“. Lematizacijai naudojama sudėtingesnė kalbinė analizė, kurioje panaudojama žodyninė bendraties forma, pavyzdžiui, paverčiant normalizuojamą žodį iš „*mylėjau*“, „*myli*“ ar „*myliu*“ į „*mylėti*“. Nors atrodo, jog šis procesas yra bereikšmis po tokenizacijos, tačiau tai vis vien gali būti itin naudinga – tokenizacija paprasčiausiai padalija tekstą į vienetus, tačiau neatlikus kamieno atrinkimo ar lematizacijos neatsižvelgiama į tai, kad skirtingos žodžių formos, pavyzdžiui „*mylintis*“, „*myli*“ arba „*mylėjo*“ gali turėti tą pačią, esminę reikšmę – „*mylėti*“. Šie du metodai sujungia žodžio variacijas į vieną formą, sumažinant žodyno dydį bei leidžiant sutelkti modelį ne į paviršinius skirtumus, tačiau į svarbesnes, teksto semantines ypatybes [21].

Svarbu paminėti, jog esą atvejų, kai nereikšmingų žodžių pašalinimas, kamieno atrinkimas ar lematizacija panaikina svarbią informaciją tolimesnei analizei. Lengvai pašalinami žodžiai ar frazės, pavyzdžiui, „*ne*“, galintys visiškai pakeisti sakinio ir sentimento reikšmę. Kaip teigia Džu ir Luo [17], renkantis pirminius duomenų apdorojimo metodus, svarbu atsižvelgti į specifinės srities aplinkybes bei tikslus, siekiant išvengti duomenų praradimo bei užtikrinant, kad modelis užfiksuos pilną tekstinės išraiškos spektrą.

Paskutinis duomenų pirminio apdorojimo žingsnis – vektorizacija. Sutvarkyti tekstiniai duomenys paverčiami į skaitinę formą, tinkančią tolimesniam kompiuteriniam apdorojimui ir klasifikavimui. Dažnai literatūroje taikyti vektorizacijos metodai – žodžių krepšelio (angl. *Bag of Words,* BoW), ir termino dažnio – atvirkštinio dokumento dažnio metodika (angl. *Term Frequency-Inverse Document Frequency,* TF-IDF). *BoW* metodui sukuriami reti vektoriai, apskaičiuojant duomenų rinkinyje pasikartojančių žodžių dažnį. *TF-IDF* metodika papildomai adaptuoja šį procesą, priskiriant atitinkamus svorius žodžiams, pagal jų pasikartojimo dažnumą duomenų rinkinyje, tokiu būdu lengviau identifikuojant informatyvesnius žodžius tolimesniam sentimentų klasifikavimui [22]. Moksliniuose tyrimuose dažnai taikoma latentinės semantinės analizės metodika, kuri įprastai naudota informacijos paieškai bei tekstinių dokumentų kategorijų atpažinimui. *LSA* pagrįsta singuliarių reikšmių dekompozicija (angl. *Singular Value Decomposition,* SVD) – *TF-IDF* metodika matrica sumažinama iki mažesnės dimensijos erdvės, kurioje slypi latentinės semantinės sąsajos tarp žodžių ir dokumentų. Tokiu būdu *LSA* metodas kompaktiškai reprezentuoja tekstą, tačiau taip pat sujungia panašius ar kontekstualiai susijusius terminus, pagerinant bendrą modelių efektyvumą, siekiant aptikti subtilius sentimentų ryšius [70]. Šios paprastesnės metodikos efektyviai veikia kartu su tradiciniais sentimentų analizės modeliais, tačiau sunkiau susidoroja su semantiniais žodžių ryšiais.

Siekiant apeiti šiuos apribojimus, sukurti žodžių įterpinių (angl. *embeddings*) metodai, kaip *Word2Vec* ar *GloVe*, panaudojantys tankius, tačiau mažų dimensijų vektorius, kuriuose įprasminami semantiniai bei kontekstiniai ryšiai tarp žodžių. Dar efektyvesni metodai, transfomeriais grįstos architektūros pagrindais sukurti sakinių ir dokumentų įterpiniai leidžia kurti kontekstinius vektorius visam sakiniui ar pastraipai. *Jina Embeddings, E5, GTE* ir kiti, *BERT* pagrindu veikiantys įterpiniai užfiksuoja ne tik semantines sąsajas, tačiau ir kalbinį kontekstą įvairiuose scenarijuose. Šie metodai vis dažniau taikomi sentimentų analizės ir teksto klasifikacijos uždaviniuose, ypač, kai dirbama su

skirtingų sričių ar kalbinių aplinkų duomenimis. Kaip teigiama naujesniuose tyrimuose, žodžių, sakinių bei dokumentų įterpinių metodikos tinkamos naudoti kartu su giliojo mokymosi (angl. *deep learning*) modeliais, tačiau paprastesni metodai, ypač *TF-IDF*, išlieka populiarūs mažesniems duomenų rinkiniams, dėl sąlyginai gero efektyvumo bei paaiškinamumo [23]

Pereinant prie pagrindinių sentimentų analizės metodų, klasifikatorių, literatūroje, kaip vieni pradinių, minimi leksikonu, (angl. *lexicon-based*) arba taisyklėmis (angl. *rule-based*) grįsti modeliai. Jiems panaudojami iš anksto apibrėžti sąrašai, kuriuose kiekvienam žodžiui priskiriamas tam tikras įvertis, pagal išreiškiamą žodžio emociją.

Leksikonu grįsti metodai įprastai skirstomi į sudarytus rankiniu būdu, automatiškai generuojamus bei hibridinius leksikonus. Rankiniu būdu sudaromi leksikonai, pavyzdžiui, *SentiWordNet* arba *AFINN* dažnai naudojami sentimentų analizėje anglų kalbai. Automatiškai generuojamiems leksikonams panaudojamos statistinės arba mašininio mokymosi metodikos, kuriomis remiantis žodžiams priskiriami sentimentų balai. Hibridiniai leksikonai sujungia šias dvi metodikas, suteikiant galimybę panaudoti jau sudarytus žodžių sąrašus kartu su mašininio mokymosi prognozėmis, taip gerinant klasifikavimo tikslumą skirtinguose kontekstuose [3]. Nors šie metodai yra efektyvūs paprastiems tekstams bei gali būti panaudojami paviršutiniškai teksto analizei, tačiau sunkiai apdorojami niuansuoti, ar kitokie sudėtingiau išreiškiami tekstiniai sentimentai apriboja šių metodikų efektyvumą [24].

Toliau literatūroje aptariamos efektyvios, mašininio mokymosi metodikos – logistinės regresijos, atraminių vektorių (angl. support vector machine, SVM), atsitiktinių miškų (angl. *Random Forest Classifier,* RF), Naivaus Bajeso (angl. *Naive Bayes,* NB), ar ekstremalaus gradiento didinimo (angl. *Extreme Gradient Boosting,* XGBoost) algoritmai.

Logistinė regresija, vienas iš klasikinių mašininio mokymosi metodų, dažnai pritaikomas binarinės klasifikacijos uždaviniuose. Nors šis metodas yra traktuojamas kaip vienas paprastesnių bei lengvai interpretuojamų, tačiau literatūroje pastebima, jog logistinė regresija yra itin konkurencingas metodas, lyginant su kitais klasifikatoriais. Siekiant išvengti persimokymo (angl. *overfitting*), bei pagerinti modelio gebėjimą apdoroti didelių matmenų duomenis, logistinė regresija dažnai pritaikoma kartu su reguliarizavimo metodų L1 (*Lasso*) bei L2 (*Ridge*) deriniu – *Elastic Net* baudomis. Šiuo metodu įgalinama požymių atranka bei modelio kompleksiškumo derinimas, kas yra svarbu atliekant sentimentų analizę [60] [68].

Kitas dažnai panaudojamas teksto klasifikavimo algoritmas, *SVM,* yra vertingas metodas, siekiant atrasti tinkamą atskyrą tarp skirtingų klasių (randant optimalią hiperplokštumą). Šis metodas plačiai pritaikomas sentimentų analizėje – kartu panaudojant atitinkamus teksto vektorizavimo metodus, efektyviai užfiksuojamas sentimentų poliariškumas. Kaip teigia Anilsagar‘as ir Syed‘as [24], dėl metodo našumo apdorojant nedidelius duomenų rinkinius bei gebėjimo generalizuoti ant nematytų duomenų, *SVM* tapo vienu esminiu metodu pradiniuose sentimentų analizės tyrinėjimo etapuose.

Atsitiktiniai miškai – ansamblinio (angl. *ensemble*) mokymosi metodas, plačiai taikomas tekstų klasifikacijos uždaviniuose ir sentimentų analizėje, dėl gebėjimo apdoroti triukšmingus bei didelių matmenų duomenis. Algoritmas sujungia daugelį nepriklausomai apmokytų sprendimo medžių – galutinės prognozės klasė parenkama daugumos balsų principu. Šia struktūra užtikrinamas efektyvesnis modelio generalizavimas bei išvengiama persimokymo rizikos, kas dažnai pasitaiko taikant pradinę šio algoritmo versiją, sprendimų medžius (angl. *Decision Tree*). Literatūroje taip pat

teigiama, jog atsitiktinių miškų klasifikatorius pasižymi stabiliais rezultatais bei funkcionalumu tais atvejais, kai duomenyse yra nereikšmingų ar triukšmingų požymių [66] [67].

Vienas iš naujesnių ansamblinio mokymosi algoritmų – *XGBoost*. Metodas panaudoja gradientų tobulinimą, tokiu būdu iteratyviai gerinant savo teikiamas prognozes, su aukštu tikslumu bei efektyvumu. Kitas, taip pat populiarus sentimentų analizės klasifikavimo metodas, dėl jo paprastumo bei efektyvumo – *NB*, bei modifikuota versija, daugianaris Naivus Bajesas (angl. *Naive Bayes Multinomial*, NBM). Dėl sąlyginės požymių nepriklausomybės prielaidos, šie modeliai greitai apmokomi bei panaudojami. *NBM* variantas yra labiau pritaikytas diskretiems požymiams, pavyzdžiui žodžių dažniams – šios modifikacijos efektyvumas išryškėja, ją panaudojant kartu su *BoW* ar *TF-IDF* metodais. Kaip ir *SVM* algoritmas, *NB* ir *NBM* metodai yra efektyvūs mažesniuose duomenų rinkiniuose [25]. Nors šie minėti modeliai yra veiksmingi naudojant mažesnius duomenų rinkinius, kuriuose yra struktūrizuoti požymiai, tačiau pateikus niuansuotus ar tam tikrai sričiai būdingus duomenų rinkinius, tradiciniai mašininio mokymosi modeliai negeba pasiekti aukštų rezultatų [26] [27].

Tradicinių metodų apribojimai paskatino naujų, giliojo mokymosi metodų, gebančių fiksuoti hierarchinius bei kontekstinius ryšius tekste, kūrybą bei adaptaciją. Sąsūkų dirbtiniai neuronų tinklai (angl. *convolutional neural networks*, CNN) buvo sukurti vaizdo analizės pritaikymui, tačiau vėliau sėkmingai pritaikyti teksto klasifikavimo uždaviniuose. Pritaikant sąsūkų naudojamus filtrus, kurie leidžia lengvai identifikuoti lokalias priklausomybes (angl. *local dependency*) tekste, pavyzdžiui, esmines frazes tam tikrose teksto vietose, taip įprasminant modelio reikšmingumą sentimentų analizės uždaviniuose. *CNN* metodas gali nesunkiai aptikti frazes, kurios yra itin svarbios nustatinėjant sentimentų poliariškumą, kaip „*labai rekomenduoju*“ arba „*nėra gerai*“. Tačiau šie modeliai negeba efektyviai išlaikyti ilgalaikių priklausomybių, leidžiančių adaptuotis prie kintančių sentimentų tekste – dėl to sumažėja šios metodikos pritaikomumas skirtinguose sentimentų analizės uždaviniuose [26].

Rekurentiniai dirbtiniai neuronų tinklai (angl. *recurrent neural networks,* RNN) buvo sukurti, siekiant apeiti šį apribojimą, apdorojant tekstą kas vieną teksto vienetą (žetoną), nuosekliai bei išlaikant kontekstinę informaciją, net per ilgesnius sakinius. Sukurti du *RNN* patobulinimai – ilgos trumpalaikės atminties modelis (angl. *Long Short-Term Memory*, LSTM) bei sklendžių rekurentinių vienetų modelis, (angl. *Gated Recurrent Units*, GRU) kurie dar labiau pagerino modelio našumą išlaikant ilgalaikes priklausomybes, t. y., ilgiau išlaikant svarbią informaciją sakiniuose, net per ilgesnius dokumentus, taip išvengiant nykstančios gradientų problemos (angl. *vanishing gradient problem*). Dėl geresnio bendro konteksto išlaikymo, ir dėl efektyviai fiksuojamų lokalių priklausomybių tyrimuose, *RNN* metodai laikomi kaip efektyvesni, lyginant su tradiciniais mašininio mokymosi metodais [27] [28].

Taip pat, atsiradus hibridiniams modeliams, sujungtos *CNN* ir *RNN* stipriosios pusės – *CNN* panaudojami, siekiant išgauti lokaliąsias savybes, o *RNN* paskirti laikinųjų priklausomybių išlaikymui. Tokiu būdu, hibridiniai modeliai geba apdoroti ir detalią informaciją, pavyzdžiui, frazes ir platesnio spektro informaciją, kaip sekų struktūras, gerinant efektyvumą sudėtingose sentimentų analizės užduotyse [26] [27].

Prieš mažiau nei dešimtmetį, 2017 metais, Vaswani ir kt. [29] „Google“ tyrėjai išleido mokslinį straipsnį, nurodant apie naujo, tobulesnio teksto analizės metodo potencialą. Transformeriais grįstos

architektūros naudoja dėmesio sutelkimo (angl. *self-attention*) metodiką. T. y., modelis geba atsižvelgti į kitus žodžius sakinyje, nustatant, kurie žodžiai yra svarbiausi, taip efektyviau perprantant prasmę ar sentimentą, bei gerinant bendrą transformerio konteksto suvokimą. Ši savybė leidžia modeliuoti ilgalaikes priklausomybes tekste be pastebimų sekvencinių apribojimų, figūruojančių *RNN* metodikose. Panaudojant lygiagretų apdirbimą, transformeriais grįstos architektūros ženkliai pagerina esamų *NLP* modelių pritaikomumą bei našumą [30]. Vienas pirmųjų, didesnių pritaikytų transfomerių modelių, *BERT*, padarė didelę įtaką sentimentų analizės efektyvumui. Tekstas apdorojamas naudojant dvikryptį metodą, skirtingai nei atlieka įprasti metodai, kurie tekstą apdirba tik iš kairės į dešinę, arba iš dešinės į kairę. Tokiu būdu visiškai įprasminamas aplinkinis žodžių kontekstas, bei geriau užfiksuojami niuansuoti tekstiniai ryšiai, taip dar labiau prisidedant prie sentimentų klasifikacijos efektyvinimo [31].

Siekiant sumažinti *BERT* modelio dydį bei panaudojamus apskaičiavimo resursus, sukurta *DistilBERT* modifikacija, išlaikanti beveik identišką pagrindinio modelio tikslumą. Panaudojant perkeltinio mokymosi metodiką, žinių distiliavimą (angl. knowledge distillation), *DistilBERT* modelis išlaiko apie 97 % *BERT* našumo, tačiau veikia 60 % greičiau, bei naudoja 40 % mažiau modelio parametrų [32].

RoBERTa modelis, sukurtas remiantis *BERT* pagrindais, patobulinus architektūrą bei apmokymo metodus, pasiekia dar geresnius rezultatus *NLP* užduotyse. Vienas reikšmingiausių patobulinimų, lyginant su *BERT*, yra sekančio sakinio nuspėjimo (angl. *next-sentence prediction task*) pašalinimas apmokant modelį. Apmokymas vyksta tik modeliui nuspėjant atskirus sakinius ar intarpus, nereikalaujant ieškoti sąryšio tarp dviejų atskirų sakinių. Taip pat, tokiu būdu modelio apmokymas orientuojasi į užmaskuotą kalbos modeliavimą (angl. *masked language modelling,* MLM) – pateikiamame modeliui tekste (sakiniuose) užmaskuojant tam tikrus žodžius, arba žetonus, siekiama prognozuoti trūkstamus žodžius pagal aplinkinį kontekstą bei prasmę. Apmokant modelį su *MLM,* panaudojamas ir dinaminis maskavimas, t. y., vis keičiant užmaskuotų žodžių vietą pateikiame tekste, užtikrinamas platesnis bei geresnis bendras modelio kontekstinis suvokimas [26] [33]. Didinant apmokymo duomenų rinkinius, *RoBERTa* modeliu pasiekiamas vis geresnis tikslumas per skirtingus *NLP* rodiklius ir sentimentų klasifikavimo uždavinius. Adaptuojant *RoBERTa* sentimentų analizei, bei panaudojant duomenų rinkinius su sudėtingomis ir dviprasmėmis sentimentų išraiškomis, modelis pranoksta tradicinę *BERT* metodiką [26].

Tačiau *RoBERTa* modelio pritaikymas daugiakalbiuose kontekstuose apribotas, nes pats modelis yra apmokytas anglų kalbos duomenimis. Kaip alternatyva daugiakalbiams tikslams, sukurta *XLM-RoBERTa* (*XLM-R*) modifikacija, naudojama būtent įvairių kalbų sentimentų analizės užduotims. Modelis buvo apmokomas ant 2,5 TB apimties *CommonCrawl* daugiakalbio duomenų rinkinio, apimančio virš 100 pasaulio kalbų, įskaitant ir lietuvių [34] [35]. Dėl šios savybės, modelis gerai pritaikomas tarptautinėse aplinkose. Kadangi vartotojų kuriamas *UGC* dažnai pateikiamas įvairiomis kalbomis, modelio galimybės identifikuoti bei analizuoti sentimentus būtent tokiuose kontekstuose ženkliai pakelia sentimentų analizės pritaikomumo lygį [31]. *XLM-R* metodu pasiekiami labai geri rezultatai sferose su ribotu skaičiumi žymėtų duomenų, adaptuojant modelį naujai užduočiai, nepateikiant pavyzdžių (angl. *zero-shot learning*), bei pateikiant kelis pavyzdžius (angl. *few-shot learning*). Panaudojant didelius kiekius daugiakalbių apmokymo duomenų, *XLM-R* efektyviai geba generalizuoti sentimentų prognozes mažai resursų turinčiose kalbose [26]. Nors transformeriais grįstų modelių efektyvumas yra aukštas, tačiau šioms metodikoms egzistuoja ir tam tikri apribojimai. Dideli skaičiavimo resursų reikalavimai transformerių modeliams, bei daugumos šių metodų priklausomybė

nuo žymėtų duomenų rinkinių sumažina prieinamumą, ypač bandant pritaikyti šiuos modelius mažesniu mastu [33].

Apžvelgti įvairūs moksliniai tyrimai, įvykdyti mažmeninės prekybos kontekstuose. Siekiant įžvelgti skirtingų sentimentų analizės metodų pritaikomumą įvairiuose scenarijuose bei kalbose, 1 lentelėje pateikiami susisteminti bei apibendrinti moksliniai darbai.

1 lentelė. Sentimentų analizės metodų pritaikymų atvejai mažmeninėje prekyboje

Autorius	**Tyrimo sritis, kalba**	**Sentimentų analizės metodai**	**Metodo privalumai**	**Metodo trūkumai**	**Rezultatai**
Ashbaugh L., Zhang Y. (2024) [50]	„Amazon“ vartotojų atsiliepimai, anglų k.	*TF-IDF*, Logistinė regresija	Efektyvumas apskaičiavimo atžvilgiu, pritaikomumas didesniems duomenų rinkiniams.	Neefektyviai fiksuojami sudetingėsni kontekstiniai šablonai.	Aukščiausias pasiektas tikslumas – 99 %
Willianto T., Supryadi S., Wibowo A. (2020) [51]	Įvairių mažmeninės prekybos svetainių atsiliepimai, indoneziečių k.	*TF-IDF*, *SVM*	Nesudėtingas metodo naudojimo paruošimas, pritaikomumas didesniems duomenų rinkiniams.	Permokymo rizika, praleidžiami kontekstiniai žodžių ryšiai.	Aukščiausias tikslumas – 85,97 %
Ashbaugh L., Zhang Y. (2024) [50]	„Amazon“ vartotojų atsiliepimai, anglų k.	*TF-IDF*, *RF*	Permokymo rizikos sumažinimas, pritaikomumas didesniems duomenų rinkiniams.	Neefektyviai užfiksuojami žodiniai ryšiai, suteikiantys kontekstinės informacijos.	Aukščiausias pasiektas tikslumas – 99 %
Ashbaugh L., Zhang Y. (2024) [50]	„Amazon“ vartotojų atsiliepimai, anglų k.	*TF-IDF*, *NB*	Lengvai pritaikomas, robastiškas nedideliuose duomenų rinkiniuose.	Prastas klasifikavimas bei tikslumas, kai žodžiai koreliuoja tarpusavyje.	Aukščiausias pasiektas tikslumas – 84 %
Wang, Z. (2025) [52]	Internetiniai mobiliųjų telefonų atsiliepimai, kinų k.	*TF-IDF*, *XGBoost*	Optimali generalizacija tarp šios srities atsiliepimų, aukštas efektyvumas.	Metodas pritaikytas būtent šiam duomenų rinkiniui, ribotas pernaudojimas kitoms sritims.	Aukščiausias pasiektas tikslumas – 88 %
Ashbaugh L., Zhang Y. (2024) [50]	„Amazon“ vartotojų atsiliepimai, anglų k.	Žodžių lygmens įterpiniai, *CNN*	Efektyviai pritaikomas dideliuose duomenų rinkiniuose, bei aptinkant žodines struktūras bei perteikiamą sentimentą.	Dėl didesnio dėmesio skiriamo lokaliam kontekstui, apribojamas platesnio masto suvokimas.	Aukščiausias pasiektas tikslumas – 93 %
Ashbaugh L., Zhang Y. (2024) [50]	„Amazon“ vartotojų atsiliepimai, anglų k.	Žodžių lygmens įterpiniai, *RNN*	Dėl naudojamų rekurentinių NN tinklų suteikiamas efektyvus, plataus masto kontekstinis suvokimas.	Didesnis apskaičiavimo resursų reikalavimai, ilgiau užtrunkantis apmokymas.	Aukščiausias pasiektas tikslumas – 98 %
Prytula M. (2024) [53]	Įvairių mažmeninės prekybos sričių ir restoranų atsiliepimai, ukrainiečių k.	*XLM-RoBERTa*	Aukštas efektyvumas dėl išankstinio apmokymo ant daugiau nei 100 kalbų.	Laiko ir apskaičiavimo resursų reikalaujantis priderinimas bei apmokymas.	Aukščiausias pasiektas tikslumas – 91,32 %

Šie apibendrinti sentimentų analizės metodai plačiai taikomi skirtinguose mažmeninės prekybos scenarijuose, įvairiomis kalbomis bei pasiteikiant tradicinius ir šiuolaikinius analizės metodus. Plati modelių bei kalbinių kontekstų įvairovė pabrėžia analitinio lankstumo svarbą, bei sentimentų analizės potencialą įvairiakalbėse ir daugiakultūrėse aplinkose. Tai yra itin aktualu vartotojų elgsenos tyrimuose, kuriuose kalbiniai bei kultūriniai niuansai gali lemti vartotojo emocinį atsaką į prekes ar paslaugas. Tokie tyrimai leidžia suprasti bendras vartotojų emocinių išraiškų tendencijas bei gali prisidėti, siekiant efektyviai adaptuoti rinkodaros ar komunikacijos strategijas skirtingiems vartotojų segmentams. Literatūroje pateikiama išvada, kad sentimentų analizė yra ne tik paviršutiniškas analitinis įrankis, tačiau ir efektyvus sprendimas, siekiant geriau suprasti vartotojų psichologiją bei elgseną [11] [52]. Kitame skyriuje aptariami tarpsritinės sentimentų analizės bei perkeltinio mokymosi principai, leidžiantys plačiau pritaikyti 1.2. skyriuje aptartas metodikas praktinėse situacijose.

1.3. Tarpsritinė sentimentų analizė

Sentimentų analizėje, sričių adaptacija yra metodas, leidžiantis efektyviai perkelti tam tikroje srityje apmokytą modelį bei jo funkcionalumą į kitą sritį. Kadangi skirtingų sričių tekstuose slypi ryškūs lingvistiniai bei semantiniai skirtumai, privalu į tai atsižvelgti, siekiant užtikrinti modelio tikslumą [36]. Pavyzdžiui, žodis „*pigus*“ gali būti traktuojamas kaip teigiamas sentimentas vartojimo prekių rinkoje, tačiau prabangos prekių segmente, šis žodis įgyja neigiamą reikšmę dėl asociacijų su prasta kokybe. Sričių adaptacijos metodikos leidžia spręsti tokius semantinius niuansus, bei užtikrina naudojamų modelių generalizacijos gebėjimus per įvairius kontekstus.

Literatūroje išskiriami trys pagrindiniai apribojimai sričių adaptacijoje – kovariatų pokyčiai (angl. *covariate shift*), žymėjimų pokyčiai (angl. *label shift*) bei koncepto pokyčiai (angl. *concept drift*). Kovariatų pokyčiai atsiranda, kai skiriasi savybių pasiskirstymas tarp šaltinio ir tikslinės srities, t. y., kai savybės, svarbios vienoje sferoje, gali būti mažiau svarbios kitoje. Žymėjimų pokyčiai kyla dėl sentimentų žymėjimų pasiskirstymo pokyčių – pavyzdžiui, tam tikros frazės vienoje srityje traktuojamos kaip neutralios, kitoje kaip teigiamos ar neigiamos. Koncepto pokyčiai pastebimi, kai ryšys tarp savybių ir žymėjimų ryšių atsiranda dėl lingvistinių ar kultūrinių skirtumų [37]. Lietuvoje, mažmeninės prekybos sektoriuje, šios problemos tampa aktualios dėl lingvistinių subtilybių ir įvairių vartotojų elgsenos niuansų.

Siekiant apeiti paminėtus apribojimus, sričių adaptacijos metodams dažniausiai naudojamas savybėmis grįstas perkėlimas (angl. *feature-based transfer*). Ši metodika identifikuoja bendras savybes skirtingose sferose, pavyzdžiui, dažnai pasikartojančius žodžius ar frazes, tokiu būdu sumažinant skirtumą tarp šaltinio ir tikslinių sričių. Taip pat panaudojami matmenų mažinimo metodai, kaip principinių komponentų analizė (angl. *principal component analysis*, PCA), leidžianti užfiksuoti latentines semantines savybes, pasikartojančias per skirtingas sritis [38]. Kita populiari metodika – atvejų svorio priskyrimas (angl. *instance weighting*), su kuria šaltinio srities pavyzdžiams priskiriami svoriai, priklausomai nuo jų panašumo į tikslinės sferos pavyzdžius. Šis metodas efektyvus, kai skirtingos sferos pasižymi nedideliais skirtumais, tačiau esant reikšmingesniems skirtumams, metodo veiksmingumas sumažėja [39].

Literatūroje pabrėžiama, jog pažangesni, giliojo mokymosi metodai ženkliai pagerina sričių adaptacijos galimybes. Jau paminėti, iš anksto apmokyti, transformeriais grįsti modeliai, kaip *BERT* pasižymi galimybėmis užfiksuoti kontekstinius bei semantinius ryšius tekste, net, kai egzistuoja

didesni skirtumai tarp sferų. Transformeriai, panaudojant dėmesio sutelkimo mechanizmus, geba nustatyti svarbiausius žodžius sakiniuose bei išlaikyti lokalias ir bendrines, visą tekstą apimančias priklausomybes [34].

Kita paminėta adaptacijos metodika literatūroje – sričių adversariniai neuronų tinklai (angl. *domain-adversarial neural networks*, DANN) – šiuo metodu efektyviai mažinamas šališkumas tarp sričių. Pasitelkiant adversarinį apmokymą, atitinkamai atskiriamos unikalios šaltinio bei tikslinių sferų savybės, tuo pačiu metu identifikuojant bendras savybes, kurios minimizuoja skirtumus tarp sričių. Naudojamas sričių klasifikatorius, siekiant modeliui leisti geriau atpažinti nekintančias savybes tarp sričių. Tokiu būdu gerinama modelio generalizacija bei užtikrinamas našumas tarp sferų su didesniais skirtumais [40].

Daugiatikslio mokymosi (angl. *multitask learning*) metodikos tuo pačiu metu apmoko vieną modelį ant kelių skirtingų uždavinių. Identifikuojant bendras savybes per kelis uždavinius, gerinami adaptacijos gebėjimai bei apeinami sričių resursų kiekio apribojimai. Metodas gali būti sėkmingai pritaikomas sentimentų klasifikacijos ar kituose panašiuose *NLP* uždaviniuose, kur apmokant modelį per skirtingus uždavinius potencialiai gerinamas efektyvumas [41].

Dar viena reikšminga tyrimų kryptis sentimentų analizėje – modelių gebėjimo generalizuoti tarp skirtingų sričių vertinimas. Tokiais atvejais pritaikoma *leave-one-domain-out* metodika, kai viena sritis paliekama kaip atskiras testavimo rinkinys, o modeliai apmokomi su likusiomis sritimis. Šis metodas leidžia įvertinti klasifikatorių atsparumą duomenų perkėlimui, bei parodo generalizavimo galimybes [69]. Tai yra aktualu, siekiant klasifikavimo sistemose užtikrinta optimalų modelių veikimą skirtingų tematikų ar kalbinių sričių tekstuose.

Jau minėti modelių adaptavimo metodai, naujai užduočiai nepateikiant žymėtų duomenų iš tikslinės srities bei pateikiant kelis žymėtų duomenų pavyzdžius (angl. *zero-shot* ir *few-shot*) leidžia pasiekti gerus rezultatus naujose srityse. *Zero-shot* modeliai, tarp kurių ir plačiai žinomas *GPT-3,* panaudoja išankstinį apmokymą, tokiu būdu leidžiant modeliui efektyviai generalizuoti ant nematytų sferų be papildomos adaptacijos [36]. Tuo tarpu *few-shot* apmokymui reikalinga adaptacija ant nedidelio skaičiaus žymėtų duomenų pavyzdžių – dėl to, tai yra taip pat vertingas pasirinkimas, analizuojant lietuvių kalbos tekstinius duomenis. Richter-Pechanski ir kt. [42] pabrėžia *few-shot* apmokymo metodų efektyvumą adaptuojant modelius pagal sričių specifiškumus. Pavyzdžiui, adaptuojant modelį šia metodika su nedideliu lietuviškų internetinių atsiliepimo duomenų rinkiniu, pagerinamas modelio efektyvumas, apeinant esminius resursų apribojimus.

Apibendrinant galima teigti, jog siekiant užtikrinti naudojamų sentimentų analizės modelių lankstumą bei tikslumą, reikalingos atitinkamos sričių adaptacijos metodikos. Literatūroje minimi tradiciniai savybių perkėlimo bei atvejų svorio priskyrimo metodai išlieka svarbūs praktiniame pritaikyme. Tačiau iškilę nauji, giliojo mokymosi metodai kaip *DANN*, bei transformeriais grįsti modeliai leidžia padidinti sričių adaptacijos galimybes. Mažai resursų turinčioms kalboms pritaikomi adaptacijos metodai kaip *zero-shot, few-shot,* daugiatikslis mokymasis bei *leave-one-domain-out,* gali būti sėkmingai pritaikomi Lietuvos mažmeninės prekybos kontekste, siekiant išgauti vertingų įžvalgų iš skirtingų vartotojų atsiliepimų sričių.

1.4. Sentimentų analizė Lietuvos mažmeninės prekybos kontekste

Vykdant sentimentų analizę Lietuvos mastu, susiduriama su unikaliais suvaržymais, giliai įsišaknijusiais į šalies kalbinius bei kultūrinius niuansus. Lietuvių kalba lingvistikoje, viso pasaulio mastu, laikoma kaip viena morfologiškai turtingiausių, dėl savo išlaikytų archajiškų savybių, neišlikusių daugumoje indoeuropiečių kalbų. Taip pat ir dėl sintaksės sudėtingumo, lietuvių kalba parašyto teksto tyryba tampa iššūkiu daugumai sentimentų klasifikavimo metodų [43].

Viena iš esminių problemų lietuviškoje sentimentų analizėje – ribotas žymėtų duomenų rinkinių kiekis. Jau egzistuojantys lietuvių kalbos duomenų rinkiniai yra sąlyginai nedideli ir dažnu atveju netvarkingi, bei turintys nemažai triukšmo. Dėl to iškyla apribojimai siekiant efektyviai apmokyti modelius, kurių našumas priklauso nuo pateikiamų duomenų. Pavyzdžiui, pagal literatūroje randamus atliktus tyrimus teigiama, jog klasikiniai mašininio mokymosi modeliai, t. y., *NB* ar *SVM* nėra itin efektyvūs lietuvių kalbos uždaviniams. Algoritmus apmokant su esamais lietuvių kalbos duomenų rinkiniais jų tikslumas dažniausiai neviršija 80 % [1].

Dar vienas iššūkis egzistuoja ties specializuotų, vienakalbių, sukurtų būtent lietuvių kalbai, modelių trūkumu. Dabartiniai didieji kalbos modeliai (angl. *large language models,* LLM), kaip *BERT* ar *T5* yra pritaikyti sentimentų analizės tikslams lietuvių kalboje. Tačiau dažnu atveju, *LLM* bazinių versijų modelių rezultatai nėra optimalūs – *DistilBERT*, *ByT5*, *XLM-RoBERTa* ir kitos adaptuotos versijos sąlyginai išsprendžia šią problemą [44]. Tolesnei sentimentų analizės pažangai reikalingi vienakalbiai *LLM*, iš anksto apmokyti būtent lietuvių kalba.

Šiandieninėje literatūroje randamas ne itin didelis kiekis reikšmingų tyrimų, atliekamų, siekiant gilinti turimas žinias bei galimybes sentimentų analizėje lietuvių kalbai. Galima teigti, kad ši sritis yra vis dar nepilnai ištyrinėta, pagrinde dėl esamų kalbinių bei išteklių apribojimų. Vieni iš pradinių, senesnių tyrimų sentimentų analizei lietuvių kalbai naudoja įprastus mašininio mokymosi algoritmus, kaip *NB* ar *SVM*, vėliau pradėtos naudoti giliojo mokymosi metodikos, *CNN* ir *RNN* bei jų pritaikymai. Naujesni tyrimai panaudoja modifikuotus *LLM* modelius siekiant optimalių rezultatų. Žemiau pateikiamoje 2 lentelėje apžvelgiami bei susisteminami sentimentų analizės lietuvių kalboje moksliniai tyrimai.

2 lentelė. Sentimentų analizės pritaikymo lietuviškam tekstui mokslinių tyrimų apibendrinimas

Autorius	**Tyrimo sritis**	**Vektorizavimo metodai**	**Mašininio mokymosi metodai**	**Duomenų rinkinys**	**Rezultatai**
Kapočiūtė-Dzikienė J., Krupavičius A., ir Krilavičius T. (2013) [20]	Internetinių portalų komentarai	*BoW*, žodžių n-gramos, simbolių n-gramos.	*NBM* ir *SVM*.	4500 internetinių komentarų iš naujienų portalo „Lietuvos rytas“ – žymėti rankiniu būdu.	Aukščiausias tikslumas – 67,9 % su *NBM*.

Autorius	**Tyrimo sritis**	**Vektorizavimo metodai**	**Mašininio mokymosi metodai**	**Duomenų rinkinys**	**Rezultatai**
Daugėla K. (2018) [49]	E-komercijos internetinių portalų atsiliepimai	*TDM*, *one-hot* koderis (angl. *one-hot encoder*).	*DT*, *RF*, *SVM*, *NN* su įterpiniais, *NN* su *LSTM* sluoksniu, *NN* su *BiLSTM* sluoksniu.	Virš 16 000 internetinių atsiliepimų, iš kurių atsitiktinai atrinkti 4000 žymėti atsiliepimai.	Aukščiausias tikslumas – 76,6 % su *NN* (įterpiniai).
Morkūnaitė, L. (2019) [46]	E-komercijos internetinių portalų atsiliepimai	*BoW*, *TF-IDF*, *LSI*, *LDA*, *RP*, *Doc2Vec*, *Sent2Vec*, *BERT*.	Logistinė regresija, *RF*, *SVM*, *XGBoost*.	Virš 18 000 internetinių atsiliepimų, 66 % iš atsiliepimų svetainės, 34 % iš „Facebook“ verslo paskyrų – žymėti rankiniu būdu.	Aukščiausias tikslumas – 91,37 %, pasiektas su *XGBoost*.
Kapočiūtė-Dzikienė J., Damaševičius R., ir Woźniak, M. (2019) [1]	Internetinių portalų komentarai	*Word2Vec* (*CboW* su neigiamų imčių atrinkimu), *FastText*.	*NBM*, *SVM*, *LSTM* ir *CNN*.	10 570 internetinių komentarų iš naujienų portalo „Lietuvos rytas“ – žymėti rankiniu būdu.	Aukščiausias tikslumas – 73,5 % su *NBM*.
Petkevičius, M., Vitkutė-Adžgauskienė D., ir Amilevičius D. (2020) [2]	Socialinių tinklų vartotojų atsiliepimai	*FastText*, *Keras* įterpiniai.	*CNN* – įprasta ir kompleksiškesnės versijos.	Bazinis duomenų rinkinys – 13 000 sakinių iš socialinių tinklų atsiliepimų apie įvairias sritis – žymėti rankiniu būdu. Augmentuotas duomenų rinkinys, praplėstas naudojant *LitWordNet* – 250 000 žodžių.	Sudėtingesniu *CNN* pasiektas aukščiausias tikslumas – 94 % aspektų klasifikavimui, 93 % sentimentų klasifikavimui.
Štrimaitis R., Stefanovič P., Ramanauskaitė S., ir Slotkienė A. (2021) [45]	Finansinių naujienų tekstai	*BoW*, žodžių įterpiniai.	*NBM*, *SVM*, *LSTM*.	10 375 finansinių tekstų iš 4 skirtingų naujienų portalų – žymėti rankiniu būdu.	Aukščiausias tikslumas – 71,1 % su *NBM*.
Vileikytė B., Lukoševičius M., ir Stankevičius L. (2024) [44]	Internetinių portalų atsiliepimai	*WordPiece* (*BERT*), *SentencePiece* (*T5*).	*DistilBERT* ir *ByT5*.	123 604 atsiliepimai – žymėti naudojant 5 žvaigždučių vertinimo sistemą, rankiniu būdu.	Aukščiausias tikslumas – 67,41 % su *DistilBERT*.

Pagal literatūros analizėje apžvelgiamus tyrimus ir teoriją, atskleidžiamas potencialas tarpsritinės sentimentų analizės darbams lietuvių kalboje. Didėjantis įvairių platformų, kaip „pigu.lt“, „atsiliepimai.lt“ ir „Google Maps“ vartotojų atsiliepimų skaičius parodo, jog moksliniuose tyrimuose ši niša gali būti ir toliau nagrinėjama.

Literatūros analize patvirtinama, jog tarpsritinė sentimentų analizė Lietuvoje, mažmeninės prekybos sektoriuje nėra plačiai išnagrinėta sritis. Egzistuoja atlikti įvairūs sentimentų analizės tyrimai, tačiau pagrinde sutelkiamas dėmesys į esminę sentimentų detekciją, neišskiriant analizės modelių pernaudojimo skirtingose srityse, tačiau atliktas naujausias rastas mokslinis tyrimas panaudoja

pažangius, transformeriais grįstus modelius, pritaikant juos internetinių atsiliepimų analizei. Pabrėžiama, kad šiuo moksliniu darbu bus prisidedama prie atliktų tyrimų, panaudojant pažangius analizės metodus, bei pritaikant juos skirtingose mažmeninės prekybos srityse Lietuvoje. Papildant ir pernaudojant lietuvių kalbos vartotojų internetinių atsiliepimų iš „evertink.lt“ ir „Facebook“ platformų duomenų rinkinį, siekiama atlikti empirinį tyrimą, kuriame bus įvertinamas pasirinktų modelių veiksmingumas analizuojant sentimentus tarp skirtingų mažmeninės prekybos sričių. Taip pat bus vertinami tyrimo rezultatai ir pateikiama jų interpretacija, bei tolimesnės rekomendacijos. Siekiama išgauti įžvalgas, kaip tarpsritinės sentimentų analizės sprendimus galima pritaikyti verslo praktikoje, norint pagerinti vartotojų elgesio įžvalgas ir sprendimų priėmimą Lietuvos mažmeninės prekybos kontekste.

2. Tyrimo metodai

Šioje darbo dalyje aprašomos tolimesniame tyrime naudojamos natūralios kalbos apdorojimo metodikos, įskaitant pasitelktus vektorizavimo metodus: *LSA, Jina Embeddings v3, Multilingual-E5-large-instruct, gte-multilingual-base* ir *xlm-roberta-base*. Taip pat aprašomi pasirinkti mašininio mokymosi algoritmai: logistinė regresija, *SVM, RF* bei *XLM-RoBERTa*, kaip atskira klasifikavimo metodika. Pateikiamos pasirinktos metrikos tyrimo rezultatams įvertinti: ROC AUC ir PRC AUC kreivės, Cohen Kappa koeficientas, F1 matas bei tikslumo, preciziškumo ir atkūriamumo įverčiai.

2.1. Duomenų vektorizavimo metodai

Siekiant turimus tekstinius duomenis pritaikyti sentimentų analizei, juos reikia paversti į atitinkamas skaitines reprezentacijas, t. y., vektorius, tinkamus mašininio mokymosi algoritmų apmokymui. Konvertuojant nestruktūrizuotą natūralios kalbos tekstą į vektorių erdvę, siekiama išlaikyti kuo daugiau semantinės bei sintaktinės teksto informacijos, reikalingos klasifikacijos modeliams efektyviai prognozuoti sentimentų klases. 2.1.1. poskyryje apžvelgiami tolimesniuose darbo etapuose taikyti vektorizacijos metodai bei paaiškinamas jų veikimo principas.

2.1.1. Latentinė semantinė analizė

Latentinė semantinė analizė yra matematinis metodas, naudojamas *NLP* uždaviniuose. *LSA* yra tinkamas klasifikacijos optimizavimo, dimensionalumo sumažinimo bei tematinės analizės tikslams, transformuojant tekstinius duomenis į mažesnės dimensijos latentinę erdvę.

Prieš pritaikant *LSA*, procesas pradedamas su *TF-IDF* vektorizacija, kuria tekstiniai duomenys konvertuojami į didesnės dimensijos retą matricą. Tokiu būdu atvaizduojama analizuojamų terminų svarba dokumente, priklausomai nuo jų dažnio visoje kalbinėje bazėje. Analizuojamiems terminams suteikiami atitinkami svoriai (svarbumas), priklausomai nuo termino dažnumo ir išskirtinumo tekstuose. Dažnai pasikartojantiems, tačiau mažiau reikšmingiems terminams, sumažinamas svoris vektorizacijos procese, kol efektyviau tekstus atskiriantiems terminams svoris yra padidinamas.

Matematiškai, *TF-IDF* svorių paskyrimas apskaičiuojamas, sujungiant du šio metodo komponentus. Termino dažniu (*TF*) apskaičiuojama, kaip dažnai terminas atsiranda analizuojamame dokumente. Skaičius normalizuojamas, dalinant iš viso terminų skaičiaus dokumente. *TF*, terminui t, esančiame dokumente d, išvedama formulė:

$$tf_{t,d} = \frac{f_{t,d}}{n_d}; \quad (1)$$

čia $f_{t,d}$ – terminų dažnis t dokumente d; n_d – visas terminų skaičius dokumente d.

Atvirkštinis dokumento dažnis (*IDF*) – šiuo metodo komponentu apskaičiuojamas tam tikro termino retumas per visą kalbinę bazę bei yra atvirkščiai proporcingas dokumentams, kuriuose egzistuoja terminas. *IDF*, terminui t, išvedama formulė:

$$idf_t = \log\frac{N}{df_t}; \quad (2)$$

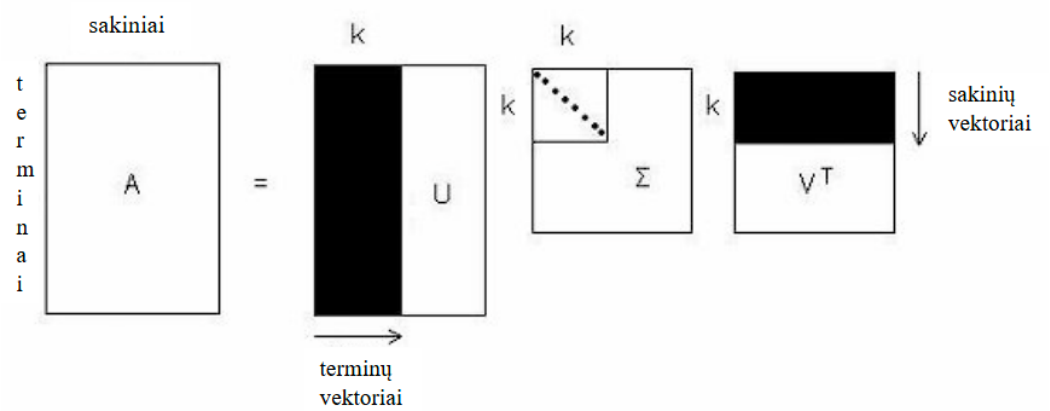
čia N – visų kalbinėje bazėje esančių dokumentų skaičius; df_t – termino dokumentų dažnis.

Galutinis *TF-IDF* svoris terminui t, dokumente d yra priskiriamas sudauginant abiejų komponentų reikšmes:

$$W_{t,d} = tf_{t,d} * idf_t; \quad (3)$$

Šia formule apskaičiuojamas bei pateikiamas termino svarbumas tam tikrame dokumente, atsižvelgiant į bendrą kalbinę bazę.

Po *TF-IDF* vektorizacijos, pritaikomas *SVD* algoritmas. *SVD* skaido *TF-IDF* matricą į tris atskiras matricas, taip sumažinant dimensionalumą, tačiau išlaikant semantinę informaciją. Atliekant šį skaidymą, (žr. 3 pav.) parenkami tik reikšmingiausi latentinių temų komponentai, atitinkantys didžiausias singuliarias reikšmes. U – terminų ir latentinių temų sąryšio matrica, Σ – diagonalioji, singuliarių reikšmių matrica, nurodanti latentinių komponentų svarbą ir V^T – dokumentų projekcija latentinių temų erdvėje.



3 pav. *SVD* metodu išgaunamos matricos (adaptuota į lietuvių kalbą pagal [53])

Tokiu būdu originali požymių erdvė sumažinama iki k dimensionalumo, išlaikant naudingiausius šablonus, užfiksuojant koreliacijas tarp terminų ir dokumentų bei pašalinant triukšmą. Kadangi sumažinamas dimensionalumas, naudojant ši metodą klasifikacijos modeliams reikalingi mažesni apskaičiavimo resursai bei įgalinama geresnė generalizacija [53] [54].

Šiame darbe *LSA* transformuoti dokumentų vektoriai panaudojami treniruojant tradicinius mašininio mokymosi klasifikacijos modelius, pritaikant juos binarinei, tarpsritinei sentimentų klasifikacijos užduočiai. Taikant *LSA*, po *TF-IDF* vektorizacijos bei *SVD* transformacijos, dimensionalumas sumažintas iki 300 komponentų. Nors *LSA* dažnai laikomas kaip vienas klasikinių metodų, tačiau tyrimuose jis išlieka aktualus kaip efektyvus palyginamasis pagrindas su šiuolaikiniais įterpinių metodais.

2.1.2. Jina modelis

Jina Embeddings v3 yra naujos kartos teksto įterpinių modelis, skirtas spręsti įvairiakalbėms *NLP* užduotims, sukurtas „Jina AI“. Transformuojant semantinę informaciją iš teksto į tankius vektorius, šis vektorizavimo metodas gali efektyviai atlikti daugumą teksto analizės uždavinių: klasifikavimą, semantinę paiešką, klasterizavimą bei semantinį tekstų atitikmenų nustatymą. Skirtingai nei tradiciniai vektorizavimo modeliai, kurie yra įprastai sukurti vienam tikslui, *Jina Embeddings v3* optimizuotas kaip daugiafunkcis įrankis, gebantis efektyviai veikti per skirtingus kalbinius bei funkcinius kontekstus. Modelio architektūra paremta *XLM-RoBERTa* transformerių modeliu su 24

paslėptais sluoksniais bei 570 milijonų parametrų, taip užtikrinant aukštą generuojamų semantinių įterpinių kokybę. *Jina Embeddings v3* palaiko iki 8192 žetonų ilgio įvestis, todėl yra gerai pritaikytas darbui su ilgesniais dokumentais ar dokumento lygmens teksto apdorojimui. Siekant pagerinti užduočių adaptaciją, modelyje integruoti žemo rango adaptavimo (angl. *Low-Rank Adaptation*, LoRA) moduliai, leidžiantys efektyviai pritaikyti šiuos teksto įterpinius specifinėms užduotims su minimaliomis apskaičiavimo sąnaudomis. Taip pat pritaikomas *Matryoshka Representation Learning* – šiuo metodu įgalinamas sklandus įterpinių dimensionalumo adaptavimas nuo įprasto 1024 iki 32, neprarandant semantinės kokybės [55].

Pagal *Massive Text Embedding Benchmark* (MTEB) rezultatus, Jina Embeddings v3 pasiekia aukštus rezultatus daugiakalbių įterpinių vertinimuose, lyginant su kitais modeliais (*Multilingual-E5-large-instruct*; *gte-multilingual base*). Tai įrodo *Jina Embeddings v3* modelio efektyvumą įvairių kalbų bei *NLP* užduočių scenarijuose [55].

Šiame darbe *Jina Embeddings v3* įterpinių modelis panaudotas generuojant 1024 dimensionalumo semantines reprezentacijas lietuvių kalbos tekstiniam duomenų rinkiniui. Šie įterpiniai taikyti kaip įvestis tradiciniams mašininio mokymosi klasifikacijos modeliams. Tokiu būdu siekiama efektyviai pritaikyti šiuolaikinius įterpinių modelius skirtingų mažmeninės prekybos sričių teksto analizės užduotims, įvertinant jų gebėjimą apdoroti duomenų heterogeniškumą, lyginant su klasikinėmis, mažiau adaptuotomis reprezentacijomis.

2.1.3. E5 modelis

E5 yra pažangus daugiakalbis teksto įterpinių modelis, skirtas semantiniam teksto reprezentavimui įvairiems *NLP* uždaviniams, sukurtas „Intfloat“. Modelis pagrįstas transformerių, t. y., *XLM-RoBERTa-large* architektūra, su 12 paslėptų sluoksnių, 109 milijonais parametrų, bei yra sukurtas spręsti įvairius *NLP* uždavinius, užtikrinant aukštos kokybės įterpinių generavimą įvairiakalbėse aplinkose. *E5*, kitaip nei klasikiniai vektorizavimo metodai, generuoja tankius vektorius, optimaliai atvaizduojančius teksto semantinį turinį, todėl tinka semantinei paieškai, tekstų klasifikacijai, rekomendavimo sistemoms bei kitoms užduotims, kurioms būtinas kontekstualus teksto supratimas. *E5* modeliai sukurti taikant dviejų etapų apmokymo strategiją. Pirmiausia, atliktas prižiūrimas kontrastinis išankstinis apmokymas (angl. *contrastive pretraining*), kurio metu modelis buvo apmokytas su daugiau nei milijardu įvairiakalbių tekstinių segmentų. Kitame žingsnyje pritaikytas prižiūrimas modelio priderinimas, panaudojant žymėtus duomenis, taip gerinant modelio efektyvumą įvairiuose *NLP* uždaviniuose [56].

Šiame darbe naudojama *Multilingual-E5-large-instruct* modelio versija. Šis variantas buvo papildomai priderintas su instrukcijomis pagrįstais duomenimis. Skirtingai nei *E5* bazinė versija, šis modelis yra optimizuotas tiksliau interpretuoti specifines instrukcijas tekste, todėl pasižymi geresne generalizacija, bei gali būti pritaikomas platesniam užduočių spektrui. Nepaisant papildomo priderinimo, modelis išlaiko *XLM-RoBERTa-large* pagrindo pranašumus, įskaitant ir įvesties palaikymą iki 8192 žetonų, todėl yra tinkamas taikyti ilgesnių tekstų analizavimui. *Multilingual-E5-large-instruct* modelio efektyvumas taip pat patvirtintas *MTEB* vertinimuose, kuriuose pasiekami aukšti įverčiai, lyginant su kitais moderniais teksto įterpinių modeliais [74].

Šiame darbe, ši *E5* modelio versija pasitelkta generuojant 1024 dimensionalumo semantines teksto reprezentacijas, kurios vėliau pritaikytos binariniam sentimentų klasifikacijos uždaviniui. Taip siekta

užtikrinti aukštą semantinio tikslumo lygį, analizuojant atsiliepimų duomenis per skirtingas mažmeninės prekybos sritis.

2.1.4. GTE modelis

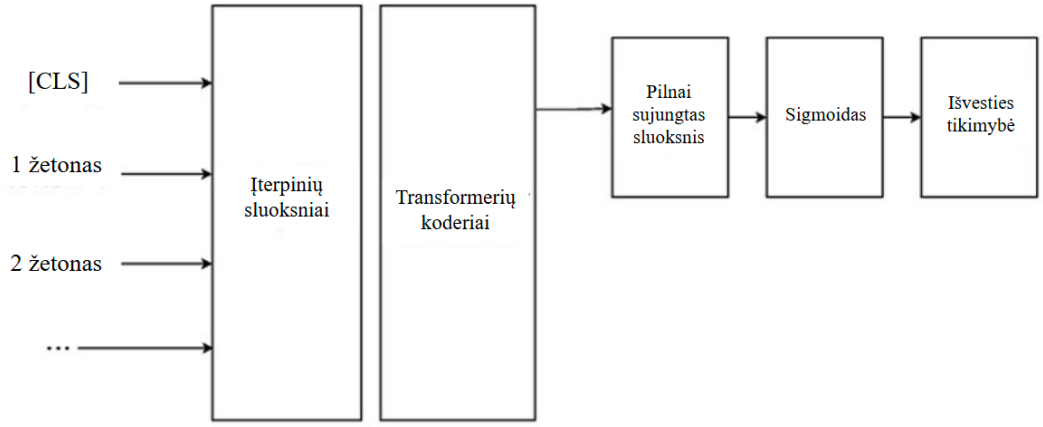
GTE (angl. *General Text Embedding*) yra pažangus teksto įterpinių modelis, sukurtas „Alibaba DAMO Academy“, siekiant efektyvių semantinių teksto reprezentacijų įvairioms *NLP* užduotims. *GTE* modeliai, kaip ir kiti šiuolaikiniai vektorizavimo metodai, transformuoja tekstą į tankius vektorius, kurie efektyviai atspindi semantinį teksto turinį. Todėl jie taip pat yra tinkami efektyviai semantinei paieškai, klasifikacijai, klasterizavimui ir kitoms teksto analizės užduotims. Bazinė *GTE* versija, su 137 milijonais parametrų, remiasi enkoderių tipo transformerių architektūra, panašia į *BERT*, tačiau papildyta keliais esminiais patobulinimais. Modelyje, vietoje įprastų absoliučių pozicinių žymų, naudojami rotaciniai poziciniai įterpiniai (angl. *Rotary Positional Embeddings*, RoPE), kurie leidžia efektyviau išsaugoti kontekstinius ryšius tarp žodžių. Taip pat, vietoje standartinių aktyvacijos funkcijų taikomi sklendžių linijiniai vienetai (angl. *Gated Linear Units*, GLU), bei papildomas „unpadding“ mechanizmas, leidžiantis sumažinti apskaičiavimo kaštus. Šiais metodais užtikrinamas efektyvus kintančio ilgio tekstų apdorojimas [72].

Šiame darbe naudojamas *gte-multilingual-base* modelis – *GTE* architektūros daugiakalbė versija, apmokyta su daugiau nei 70 kalbų duomenimis. Modelis išlaiko esminius architektūrinius principus, tačiau yra geriau pritaikytas apdoroti heterogeniškus duomenis. Tai leidžia efektyviai dirbti tiek su dažnai pasitaikančiomis, tiek su mažiau paplitusiomis kalbomis. Modelis palaiko iki 8192 žetonų ilgio įvestį, bei generuoja 768 dimensionalumo vektorius, tinkamus dokumento, sakinio ar frazės lygmens analizei. Nors *gte-multilingual-base* nenaudoja pažangių metodų, kaip ankščiau minėtų instrukcijomis pagrįstų duomenų ar *LoRA* adapterių, tačiau pagal *MTEB* vertinimą, modelis vis vien yra konkurencingas, ypač įvairiakalbiuose scenarijuose [73]. Šiame darbe, ši *GTE* modelio versija naudota generuoti semantines teksto reprezentacijas, kurios buvo taikomos sentimentų klasifikacijai, analizuojant skirtingų mažmeninės prekybos sričių klientų atsiliepimus. Tai leido įvertinti modelio gebėjimą efektyviai apdoroti semantiškai įvairų ir kontekstualiai sudėtingą tekstinį turinį.

2.1.5. XLM-RoBERTa modelis

XLM-RoBERTa yra pažangus, daugiakalbis transformerių architektūra paremtas kalbos modelis, išvystytas „FacebookAI“. Sukurtas, remiantis *BERT* principu, tačiau yra labiau pritaikytas apdoroti didelės apimties tekstinius duomenis daugiau nei šimtui kalbų. Šio modelio pagrindą sudaro *RoBERTa* (angl. *Robustly optimized BERT approach*) modelio išplėtota versija, su 12 paslėptų sluoksnių bei 279 milijonais parametrų. *XLM-RoBERTa* apmokyta su 2,5 TB „CommonCrawl“ nežymėtų duomenų rinkiniu, siekiant pagerinti modelio universalumą ir semantinį atitikimą įvairiose kalbinėse aplinkose. Modelio architektūra leidžia generuoti aukštos kokybės kontekstualius tekstinius įterpinius, išlaikant optimalią semantinę reprezentaciją net ir mažiau paplitusioms kalboms.

XLM-RoBERTa, skirtingai nei klasikiniai vektorizavimo metodai, generuoja kontekstualizuotas žodžių ir sakinių reprezentacijas, nesudėtingai pritaikomas įvairioms *NLP* užduotims: teksto klasifikacijai, semantiniams panašumams lyginti ir kt. Modelis efektyviai atvaizduoja sudėtingus sakinių ryšius, bei leidžia identifikuoti analizuojamame tekste perteikiamus ketinimus bei sentimentą. Pateikiama *BERT* ir *XLM-RoBERTa* modelių bazinė logine schema (žr. 4 pav.).



4 pav. *BERT* ir *XLM-RoBERTa* modelių loginė schema, (adaptuota į lietuvių kalbą pagal [57])

Šiame darbe *XLM-RoBERTa* modelis pritaikytas dvejais būdais. Pirmiausia buvo pasitelkti fiksuoto ilgio teksto įterpiniai, išgauti iš galutinio kodavimo sluoksnio CLS žetono (angl. *CLS token*). Jie panaudoti kaip požymiai, apmokant klasikinius klasifikacijos modelius. Tokiu būdu siekta įvertinti *XLM-RoBERTa* generuojamų reprezentacijų tinkamumą sentimentų klasifikacijos uždaviniams tarp skirtingų mažmeninės prekybos sričių.

Antruoju metodu, *XLM-RoBERTa* modelis buvo pritaikytas kaip atskiras klasifikatorius, t. y., vietoje požymių išgavimo, modelis buvo papildomai apmokytas, panaudojant žymėtą duomenų rinkinį, bei tiesiogiai generuojant prognozes. Išbandyti modelio gebėjimai adaptuotis tarp skirtingų sričių sentimentų klasifikavimo uždavinyje, nesinaudojant tradiciniais klasifikatoriais. Tiesioginio apmokymo (angl. *end-to-end*, E2E) modeliavimas leido įvertinti visos architektūros bendrą efektyvumą ir jos jautrumą duomenų perkėlimo atveju [58].

Aukšta *XLM-RoBERTa* modeliavimo kokybė patvirtinama *MTEB* vertinimo platformoje, kurioje modelis pasižymi aukšta semantine apimtimi, ypač daugiakalbiuose kontekstuose. Šiame darbe gauti rezultatai leidžia įvertinti šio modelio konkurencingumą, lyginant su kitais vektorizavimo metodais, tiek naudojant klasikinį požymių išgavimo metodą su 768 dimensionalumo vektoriais, tiek taikant *E2E* klasifikaciją.

2.2. Klasifikavimo algoritmai

Klasifikavimo uždavinys sudaro esminę daugelio tekstinės analizės projektų dalį – tokiu būdu siekiama automatiškai priskirti dokumentus iš anksto apibrėžtoms kategorijoms. Šiame darbe klasifikavimo užduotis orientuota į binarinį sentimentų klasifikavimą. Atsižvelgiant į tai, kad tekstiniai dokumentai buvo transformuoti į fiksuoto ilgio vektorių reprezentacijas, klasifikavimo užduotis atlikta pritaikant kelis skirtingus algoritmus. Reguliarizuota logistinė regresija pasižymi optimaliomis interpretacijos galimybėmis bei modelio stabilumu. *SVM* dažnai pateikia tikslesnes prognozes, dirbant su didelių dimensijų duomenimis. *RF* pasižymi atsparumu triukšmui bei turi privalumą dėl persimokymo problemos eliminavimo. Šių metodų palyginimas (kartu su įvairiais vektorizavimo metodais) leidžia įvertinti, kurie iš jų efektyviausiai sprendžia binarinę sentimentų klasifikavimo užduotį, vertinant sentimentų analizės perkeliamumą skirtingų mažmeninės prekybos sričių scenarijuose.

2.2.1. Reguliarizuota logistinė regresija

Logistinė regresija yra plačiai taikomas statistinis modelis, optimaliai pritaikytas binariniams klasifikacijos uždaviniams. Logistinės regresijos modelis įvertina išvesties tikimybę, naudojant sigmoidinę funkciją:

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(\beta_0 + x^T \beta)}}; \quad (4)$$

čia x – požymių vektorius išgaunamas iš tekstinės imties; β – modelio koeficientų vektorius; β_0 – laisvasis narys.

Sigmoidinės funkcijos išvestis interpretuojama kaip tikimybė, kad analizuojamas tekstinis atvejis (atsiliepimas) klasifikuojamas kaip 1 (teigiamas sentimentas) arba 0 klasės (neigiamas sentimentas).

Siekiant sumažinti modelio persimokymo riziką bei pagerinti generalizaciją, pritaikoma *Elastic Net* reguliarizacija. Šis metodas sujungia du klasikinius baudų metodus: L1 (*Lasso*), kuria tam tikri koeficientai mažinami iki nulio, taip optimizuojant vektorių retumą, bei L2 (*Ridge*), kuria išvengiama didelių koeficientų verčių, mažinant jas link nulio. L1 bauda naudinga požymių atrinkimui, efektyviai filtruojant neinformatyvius žodžius, o L2 bauda gerina stabilumą bei robastiškumą koreliuotų požymių atvejais. Sujungiant šias dvi baudas, formuluojama *Elastic Net* išraiška:

$$Bauda = \lambda_1 \| \beta \|_1 + \lambda_2 \| \beta \|_2^2; \quad (5)$$

čia λ_1 ir λ_2 – neneigiami reguliarizacijos parametrai, kuriais yra nustatomas L1 ir L2 baudų įtaka modelio veikimui.

Reguliarizuotos logistinės regresijos modelio privalumas – gebėjimas vienu metu atlikti požymių atranką (L1) bei užtikrinti koeficientų stabilumą (L2). Toks modelis yra efektyviai pritaikytas tekstų klasifikavimui, ypač tada, kai tekste yra daug nereikšmingų ar koreliuotų požymių. Šiame darbe reguliarizuotos logistinės regresijos klasifikatorius pasitelktas sentimentų klasifikacijai, taikant apžvelgtus vektorizavimo metodus [59] [60]. Siekiant užtikrinti optimalų metodų veikimą, atitinkamai buvo priderinti parametrai kiekvienam vektorizavimo ir šio klasifikavimo modelių kombinacijoms, panaudojant C ir $l1\_ratio$. Parametras C apibrėžia modelio reguliarizacijos stiprumą: mažesnės reikšmės užtikrina didesnę reguliarizaciją, kas padeda išvengti permokymo, tačiau gali sumažinti modelio lankstumą. Didesnės parametro reikšmės užtikrina mažesnę reguliarizaciją ir sudėtingesnį modelį, tačiau taip padidėja permokymo rizika. Tinkamos reikšmės buvo ieškomos intervale $C \in \{0{,}01;\ 0{,}1;\ 1;\ 10\}$. Parametras $l1\_ratio$ apibrėžia L1 ir L2 baudų santykį, siekiant išlaikyti pusiausvyrą tarp efektyvios požymių atrankos ir modelio koeficientų stabilumo. Reikšmės buvo ieškomos intervale $l1\_ratio \in \{0{,}2;\ 0{,}5;\ 0{,}8\}$. Modelio parametrų derinimui pasitelktas tinklelio paieškos (angl. *grid search*) metodas, taikant 3 kartų kryžminę validacija (angl. *cross validation*), bei vertinant pagal ROC AUC metriką.

2.2.2. Atraminių vektorių klasifikatorius

SVM yra efektyvus klasifikavimo algoritmas, plačiai taikomas tekstinių duomenų analizėje. Metodo privalumas – gebėjimas efektyviai apdoroti didelio dimensionalumo duomenis. *SVM* metodu siekiama rasti optimalią hiperplokštumą, kuri maksimaliai atskiria skirtingų klasių duomenų taškus. Šiame darbe taikoma tiesinė *SVM* modifikacija, pritaikyta tekstiniams duomenims, pasižymintiems

dideliu požymių skaičiumi bei retais vektoriais. Matematiškai hiperplokštumos paieškos funkcija išreiškiama šia formule:

$$f(x) = w^T x + b\,; \tag{6}$$

čia $x \in \mathbb{R}^p$ – požymių vektorius; w^T – svorių (koeficientų) vektorius; b – poslinkio (angl. *bias*) parametras.

Klasifikavimas atliekamas atsižvelgiant į ženklą: kai $f(x) > 0$, pavyzdys priskiriamas vienai klasei, kai $f(x) < 0$, priskiriama kitai. *SVM* modeliu siekiama maksimaliai padidinti skirtumą tarp artimiausių skirtingų klasių taškų (atraminių vektorių) bei sprendimo ribos – šis procesas vadinamas paraščių (angl. *margin*) maksimizavimu. Tiesiniu *SVM* atveju, taikoma paprasta branduolio (angl. *kernel*) funkcija:

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

čia $x_i, x_j \in \mathbb{R}^p$ – požymių vektoriai.

Šia funkcija apskaičiuojama dviejų vektorių panašumas įvesties erdvėje, nepaverčiant jų į didesnės dimensijos požymių erdvę. Tokiu būdu modelis tampa efektyvesnis bei efektyviau pritaikomas teksto klasifikavimo uždaviniams, kuriuose požymių vektoriai yra didelių dimensijų. Apmokant modelį, sprendžiama optimizavimo užduotis, kurioje minimizuojama užduoties funkcija:

$$\min_{w,b} \frac{1}{2} \parallel w \parallel^2 + C \sum_{i=1}^{n} \xi_i\,; \tag{8}$$

su apribojimais:

$$y_i(w^T x_i + b) \geq 1 - \xi_i; \qquad \xi_i \geq 0; \tag{9}$$

čia $y_i \in \{-1,1\}$ – tikrosios klasių žymos; ξ_i yra nuokrypio kintamieji (angl. *slack variables*), leidžiantys tam tikrą dalį apmokymo duomenų klasifikuoti neteisingai; $C \geq 0$ – baudos (reguliarizavimo) parametras.

C parametru apibrėžiamas kompromisą tarp paraščių pločio bei klasifikavimo klaidų. Kuo C reikšmė didesnė, tuo modelis yra labiau orientuotas į tikslesnį klasifikavimą. Mažesnės C reikšmės skatina modelio bendrinimą ir prisideda prie labiau apibendrinto klasifikavimo uždavinio sprendimo [61] [62] [63].

Tiesiniam *SVM* modeliui pritaikyti tekstiniai duomenys, paversti į skaitinę formą taikant ankstesniuose poskyriuose minėtus vektorizavimo metodus. Dėl optimalaus klasių atskyrimo bei efektyvumo didelio dimensionalumo požymių erdvėse, *SVM* su tiesiniu branduoliu buvo pasirinktas kaip vienas iš metodų sentimentų klasifikacijai tarp skirtingų mažmeninės prekybos sričių. Siekiant rasti optimalų C parametrą, vykdyta tinklelio paieška intervale $C \in \{0{,}001;\ 0{,}01;\ 0{,}1;\ 1;\ 10\}$, naudojant 3 kartų kryžminę validaciją, vertinant pagal ROC AUC metriką.

2.2.3. Atsitiktinių miškų klasifikatorius

Atsitiktinių miškų klasifikatorius – ansamblinio mokymosi metodas, paremtas daugybės sprendimo medžių generavimu ir jų pateikiamų prognozių sujungimu. Kiekvienas medis formuojamas remiantis

atsitiktinai parinktais požymių ir duomenų poaibiais, taip užtikrinant individualių medžių įvairovę bei sumažinant modelio dispersiją. Galutinė prognozė nustatoma daugumos balsų principu, sujungiant individualius medžių sprendimus, sukuriant robastiškesnę bei tikslesnę galutinę įšvestį.

Apmokymo metu kiekvienas medis yra auginamas pagal *CART* (angl. *Classification and Regression Trees*) metodologiją, taikant „bootstrap“ mėginių ėmimą ir atsitiktinį požymių atrinkimą kiekviename padalijimo taške. Tokiu būdu sumažinama persimokymo rizika bei pagerinamas modelio gebėjimas apdoroti triukšmingus bei koreliuotus duomenis. Be to, ši struktūra leidžia įvertinti požymių svarbą, remiantis jų įtaka sprendimų priėmimui medžiuose, modelis išlieka efektyvus bei pritaikomas prie duomenų rinkinio dydžio ar požymių skaičiaus kitimo. [64] [66].

Šiame darbe atsitiktinių miškų klasifikatorius buvo taikomas sentimentų klasifikacijos uždaviniui, pasitelkiant ankstesniuose poskyriuose minėtus metodus, vektorizuojant tekstinius dokumentus. Šis klasifikacijos algoritmas pasirinktas dėl savo atsparumo duomenų kintamumui, bei optimalaus generalizacijos ir tikslumo balanso, reikalingo tarpsritinei sentimentų analizei. Taip pat atliktas parametrų derinimas, parenkant tinkamą max_depth reikšmę – kiekvieno atskiro medžio gylį apmokymo procese, siekiant rasti tinkamą pusiausvyrą tarp modelio kompleksiškumo ir generalizavimo gebėjimų, bei mažinant persimokymo riziką. Reikšmės buvo ieškomos intervale $max\_depth \in \{none, 50, 100\}$. Kitu parametru, $max\_features$, nusakoma, kiek požymių kiekvienas medis atsirenka atskiruose padalijimo taškuose, siekiant optimizuoti balansą tarp medžių įvairovės ir apmokymo kokybės. Šio parametro reikšmės buvo parinktos remiantis viso požymių skaičiaus kvadratine šaknimi, t. y., $max\_features \in \{0{,}5 * \sqrt{n};\ \sqrt{n};\ 2 * \sqrt{n}\}$, kur n yra požymių skaičius. Atsitiktinių miškų klasifikatoriui parinktas pastovus, $n\_estimators = 250$, parametro įvertis, nurodantis medžių skaičių modelyje. Esant didesniam kiekiui medžių, potencialiai pagerinamas tikslumas, tačiau dėl to gali didėti apskaičiavimo laikas. Parametrų paieškai naudota tinklelio paieška, kartu su 3 kartų kryžmine validacija, vertinant pagal ROC AUC metriką.

2.3. Tarpsritinės sentimentų analizės eksperimento metodika

Šiame darbe siekiama įvertinti teksto vektorizacijos metodų ir klasifikatorių gebėjimą generalizuoti tarp skirtingų mažmeninės prekybos sričių atsiliepimų. Šiai užduočiai įvykdyti buvo taikoma kryžminė sričių (angl. *cross-domain*) vertinimo sistema. Eksperimentuose naudotos septynios sritys, atitinkančios skirtingas mažmeninės prekybos parduotuvių kategorijas, kurios buvo sukurtos duomenų paruošimo etape: *E-Niche, E-Marketplace, E-Tech, Groceries, Beauty, Clothing* ir *Other*. Išskyrus *Other*, kiekviena sritis buvo paeiliui naudojama kaip apmokymo sritis. Modelis buvo testuojamas visose kitose srityse, įskaitant ir pačią mokymo sritį. Tokiu būdu galima vertinti tarpsritinę modelių generalizaciją bei jų robastiškumą atpažįstant tos pačios srities tekstų sentimentus.

Iš kiekvienos pasirinktos apmokymo srities 80 % duomenų panaudota modelio apmokymui. Duomenų dalis atrinkta pasitelkus stratifikuoto atsitiktinio padalijimo (angl. *Stratified Shuffle Split*) metodą. Tokiu būdu siekta išlaikyti sričių klasių balansą ir sumažinti permokymo riziką dėl skirtingų sričių atsiliepimų kiekio. Kiekvienas eksperimentinis scenarijus buvo vykdomas atskirai, naudojant skirtingas teksto vektorizavimo bei klasifikatorių kombinacijas. Prieš vykdant tarpsritinius eksperimentus, kombinacijoms atskirai buvo vykdomas parametrų derinimas, išlaikant juos vienodus visuose tolimesniuose eksperimentų etapuose.

Eksperimente *Other* sritis buvo įtraukta tik kaip testavimo sritis. Toks pasirinkimas pagrįstas tuo, kad šios srities įmonės (ir vartotojų atsiliepimai) yra ribotai susiję su šiame darbe akcentuojama

mažmeninės prekybos tematika. Paliekant *Other* kaip modeliams nežinomą testavimo sritį, galima įvertinti modelių atsparumą duomenų perkeliamumo atveju, bei nustatyti jų taikymo galimybes ir apribojimus [39] [69].

Modelių vertinimas buvo atliekamas remiantis šiomis rezultatų metrikomis: ROC AUC, PRC AUC, Cohen Kappa koeficientu, F1 mato įverčiu, sumaišymo matricomis bei tikslumu, preciziškumu ir atkūriamumu. Rezultatai buvo kaupiami atskirai kiekvienai teksto vektorizacijos ir klasifikatoriaus kombinacijai.

2.4. Tikslumo vertinimas

Klasifikavimo uždaviniuose, mašininio mokymosi modelių efektyvumo vertinimui būtina remtis kiekybiniais tikslumo įverčiais, leidžiančiais objektyviai įvertinti modelių prognozavimo kokybę. Šiame darbe taikomi keli vertinimo matavimo metodai, apimantys tiek bendrai pripažintas, tiek klasifikavimo uždaviniams aktualias metrikas. Naudojamos metrikos leidžia įvertinti modelių gebėjimus tiksliai prognozuoti klasę, modelio jautrumą disbalansui, prognozavimo stabilumą įvairiose duomenų srityse, bei gebėjimą atitinkamai reaguoti į ribinius ar neapibrėžtus atvejus. Šiame poskyryje aptariamos penkios pagrindinės metrikos, naudotos vertinant teksto vektorizacijos ir klasifikatoriaus kombinacijų efektyvumą: ROC kreivės plotas (ROC AUC), preciziškumo-atkūriamumo kreivės plotas (PRC AUC), Cohen Kappa koeficientas, F1 matas, bei klasikiniai klasifikavimo įverčiai – tikslumas (angl. *accuracy*), preciziškumas (angl. *precision*) ir atkūriamumas (angl. *recall*). Darbe ROC AUC ir PRC AUC metrikos buvo pasirinktos kaip pagrindinės vertinimo priemonės, dėl tinkamumo vertinti modelių veikimą esant sentimentų klasių disbalansui, kas yra aktualu eksperimente naudojamame duomenų rinkinyje.

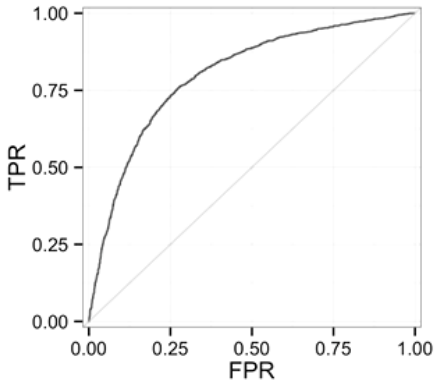
Vienas iš universaliausių ir plačiai taikomų klasifikavimo modelių vertinimo būdų yra ROC (angl. *Receiver Operating Characteristic*) kreivė, bei jos integralas – AUC (angl. *Area Under the Curve*). ROC kreivė atvaizduoja modelio klasifikavimo gebėjimą įvairiuose slenksčių lygiuose. Kreivėje pateikiamas ryšys tarp modelio teigiamai identifikuotų atvejų (angl. *True Positive Rate,* TPR), t. y., jautrumo, ir klaidingai teigiamų atvejų dalies (angl. *False Positive Rate,* FPR). Šios dvi reikšmės apskaičiuojamos pagal šias formules:

$$TPR = \frac{TP}{TP + FN}; \tag{10}$$

$$FPR = \frac{FP}{FP + TN}; \tag{11}$$

čia TP (angl. *True Positive*) – teisingai identifikuoti teigiami atvejai; FN (angl. *False Negative*) – neteisingai identifikuoti neigiami atvejai; FP (angl. *False Positive*) – neteisingai identifikuoti teigiami atvejai; TN (angl. *True Negative*) – teisingai identifikuoti neigiami atvejai.

Kiekviename klasifikavimo slenksčio taške modelis pateikia skirtingą *TPR* ir *FPR* reikšmių porą, iš kurių sudaryta ROC kreivė. Kreivės plotas, esantis po ROC kreive, tačiau virš atsitiktinės spėjimo tiesės (kuri eina įstrižai, į dešinį kampą), yra AUC matas, matuojamas skalėje nuo 0 iki 1. Aukščiausia reikšmė – 1, kuri reiškia, jog klasifikatorius yra idealus, bei visus atvejus atskiria be klaidų; 0,5 reikšmė rodo atsitiktinį spėjimą, t. y., modelio klasifikavimo gebėjimai nėra geresni nei atsitiktinis spėjimas (žr. 5 pav.). Reikšmės žemesnės nei 0,5 parodo, jog modelis klasifikuoja prasčiau nei atsitiktiniai spėjimai.

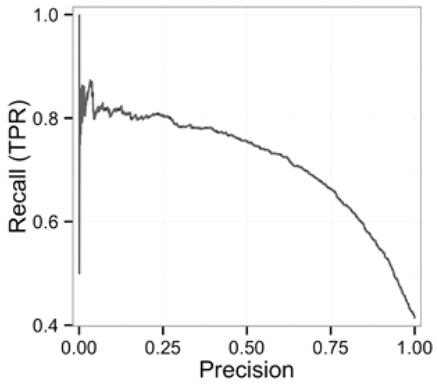


5 pav. ROC kreivė [65]

Šis vertinimo metodas naudingas, kai duomenų rinkinyje egzistuoja klasių disbalansas. ROC AUC vertina modelio gebėjimą ranginiu būdu atskirti teigiamus ir neigiamus atvejus, nepriklausomai nuo pasirinkto klasifikavimo slenksčio, todėl ši metrika tampa universaliu matu, vertinant binarinių klasifikatorių efektyvumą įvairiuose kontekstuose.

Kita esminė binarinių klasifikacijos modelių vertinimo metrika yra preciziškumo-atkūriamumo kreivės plotas, PRC (angl. *Precision-Recall Curve*) AUC. Ji taip pat yra naudinga, vertinant modelio veikimą esant duomenių klasių disbalansui, realiuose duomenų rinkiniuose. Nors PRC AUC dažnai taikomas esant duomenų disbalansui, kai dominuoja neigiama klasė, tačiau metrika yra informatyvi ir atvirkštinėje situacijoje, kai dominuoja teigiama klasė, kaip ir šiame darbe. Dėl to PRC AUC metrika leidžia įvertinti, ar modelis nėra pernelyg optimistiškas, t. y., kai modelis per dažnai prognozuoja teigiamas klases, dėl ko sumažėja preciziškumas.

Vertinimui naudojamas PRC plotas po kreive (AUC), kurio reikšmė taip pat kaip ir ROC, svyruoja nuo 0 iki 1. Kuo didesnė PRC AUC reikšmė, tuo efektyviau modelis geba aptikti teigiamos klasės pavyzdžius, ir tuo pačiu metu išvengti klaidingų teigiamų prognozių. Idealu, kai PRC AUC reikšmė yra artima 1 – tai parodo, kad modelis pasiekia aukštą preciziškumą ir tikslumą visame slenksčių intervale (žr. 6 pav.).



6 pav. PR kreivė [65]

Pateikiamos preciziškumo ir atkūriamumo apskaičiavimo formulės, naudojamos PR kreivės sudarymui:

$$Preciziškumas = \frac{TP}{TP + FP}; \quad (12)$$

$$Atkūriamumas = \frac{TP}{TP + FN}; \quad (13)$$

PR kreivė atvaizduoja ryšį tarp preciziškumo ir atkūriamumo kintant klasifikavimo slenksčiui. Kiekvienas kreivės taškas atspindi konkretų modelio veikimo kompromisą tarp šių dviejų dydžių – didesnis atkūriamumas mažina preciziškumą ir atvirkščiai. PRC AUC plotas atspindi modelio gebėjimą išlaikyti aukštą preciziškumą, teisingai prognozuojant didžiąją dalį teigiamų atvejų. Duomenų rinkiniuose, kuriuose yra daug teigiamų klasių, metrika leidžia identifikuoti, ar modelis yra linkęs neteisingai identifikuoti neigiamus atvejus dėl disbalanso – ši vertinimo metrika naudinga, vertinant skirtingų modelių balansą tarp preciziškumo bei atkūriamumo.

Kita vertinimo metrika, Cohen Kappa koeficientas, taikomas klasifikavimo uždaviniuose, siekant įvertinti sutapimo lygį tarp klasifikatoriaus prognozių ir tikrųjų duomenų žymų, atsižvelgiant į atsitiktinio sutapimo tikimybę. Šis rodiklis iš pradžių sukurtas ir naudotas dviejų žmogiškųjų vertintojų duomenų žymėjimų palyginimui. Tačiau mašininio mokymosi kontekste, metrika pritaikyta lyginant modelio prognozes bei teisingas teksto žymas. Tokia adaptacija leidžia taikyti šį koeficientą klasifikavimo uždaviniams, ypač nesubalansuotai pasiskirsčiusiems duomenų rinkiniams. Kappa koeficientas apskaičiuojamas pagal šią formulę:

$$\kappa = \frac{p_0 - p_e}{1 - p_e}; \quad (14)$$

čia p_0 – pastebimas sutapimas tarp prognozių ir tikrųjų žymų; p_e – tikėtinas atsitiktinis sutapimas, apskaičiuotas pagal prognozuojamų ir faktinių žymų pasiskirstymą.

Kappos koeficiento reikšmė 1 rodo visišką sutapimą; 0 – atsitiktinį sutapimą; o neigiamos reikšmės parodo sisteminį nesutapimą tarp prognozių ir žymų. Tarpsritinėje sentimentų analizėje Kappa koeficientas pateikia informatyvesnį bei statistiškai patikimesnį modelio vertinimą, lyginant su paprasta modelio tikslumo metrika. Kappa koeficientas leidžia įvertinti, ar modelis geba generalizuoti tarp skirtingų sričių, neapsiribojant dominuojančios klasės dažniu.

Kita metrika, F1 matas, gali būti taip pat naudingas, dirbant su nesubalansuotų klasių duomenimis. Šiame rodiklyje sujungiamas preciziškumas ir atkūriamumas, apskaičiuojant jų harmoninį vidurkį. Taip užtikrinama pusiausvyra tarp neteisingai identifikuojamų teigiamų ir neigiamų prognozių. F1 reikšmė apskaičiuojama pagal šią formulę:

$$F1 = 2 * \frac{preciziškumas * atkūriamumas}{preciziškumas + atkūriamumas}; \quad (15)$$

Rodiklio įvertis svyruoja nuo 0 iki 1. Reikšmė lygi 1 rodo idealią klasifikaciją, kol reikšmės artimos 0 rodo prastą balansą tarp preciziškumo ir atkūriamumo. F1 rodiklis yra naudingas vertinant klasifikatoriaus efektyvumą tais atvejais, kai svarbu pasiekti kompromisą tarp šių dviejų esminių klasifikavimo aspektų.

Įprastai, binarinėms klasifikavimo užduotims įvertinti taikomos trys pagrindinės metrikos – tikslumas, preciziškumas ir atkūriamumas. Šios metrikos padeda įvertinti klasifikatoriaus gebėjimus, atsižvelgiant į teisingai bei neteisingai identifikuotus atvejus.

Tikslumas apibrėžiamas kaip visų teisingų prognozių (teigiamų ir neigiamų) dalis, visų modelio prognozių atžvilgiu. Tai viena dažniausiai taikomų ir lengviausiai interpretuojamų vertinimo priemonių, tačiau ji gali klaidinti esant reikšmingam klasių disbalansui. Preciziškumas nurodo, kokia dalis prognozuotų teigiamų atvejų iš tikrųjų buvo teisingi. Atkūriamumas atspindi, kiek iš tikrųjų teigiamų atvejų modelis sugeba teisingai identifikuoti.

Siekiant vizualiai įvertinti išbandomų metodų gebėjimus atpažinti teigiamas ir neigiamas klases, naudojamos sumaišymo matricos. Šios matricos yra aktualios sentimentų analizės kontekste, ypač klasių disbalanso atvejais. Lentelėje (arba iliustracijoje) pateikiamos modelio prognozės TN, FP, FN ir TP atvejais, leidžia tiksliai įvertinti pasirinkto klasifikacijos metodo galimybes. Sumaišymo matricose taip pat pateikiami preciziškumo ir atkūriamumo rodikliai, kas leidžia įvertinti modelio veikimą kiekvienos klasės atžvilgiu.

3. Tyrimų rezultatai

Šioje darbo dalyje praktiškai pritaikomos antroje baigiamojo projekto dalyje aprašytos vektorizavimo bei klasifikavimo metodikos. Siekiama įvertinti naudojamų metodų efektyvumą tarpsritinės sentimentų analizės Lietuvos mažmeninės prekybos atveju, naudojamasi atitinkamai paruoštu lietuvių kalbos internetinių atsiliepimų duomenų rinkiniu. Eksperimentams įvykdyti buvo pasinaudota „*AI KTU Notebook*“ platforma, suteikianti prieigą prie didelio našumo *NVIDIA* H100 NVL GPU, taip siekiant užtikrinti efektyvius rezultatų apskaičiavimus su visais naudojamais vektorizavimo bei klasifikavimo metodais. Eksperimentams atlikti naudota Python programavimo kalba.

3.1. Pasirinkto duomenų rinkinio paruošimas

Šiame darbe pernaudojamas duomenų rinkinys su 18539 internetiniais atsiliepimais iš įvairių Lietuvoje veikiančių verslų. Pirminis duomenų rinkinys sudarytas iš 66 proc. atsiliepimų surinktų iš „evertink.lt“ nuomonės sklaidos tinklapio, bei 34 proc. atsiliepimų, surinktų iš „Facebook“ verslo paskyrų atsiliepimų skilties. Atsiliepimai surinkti 2011-2018 metų laikotarpiui. Duomenų rinkinyje sukurti šie stulpeliai, panaudoti tyrimams: atsiliepimo šaltinis, įmonė, paskelbimo data, vartotojo parašytas tekstas, prie atsiliepimo paliekamas vartotojo įvertinimas $\in \{1, \ldots, 5\}$, bei rankiniu būdu įvykdytas klasių žymėjimas, kur 1 žymimi teigiami, 0 žymimi neigiami atsiliepimai [46].

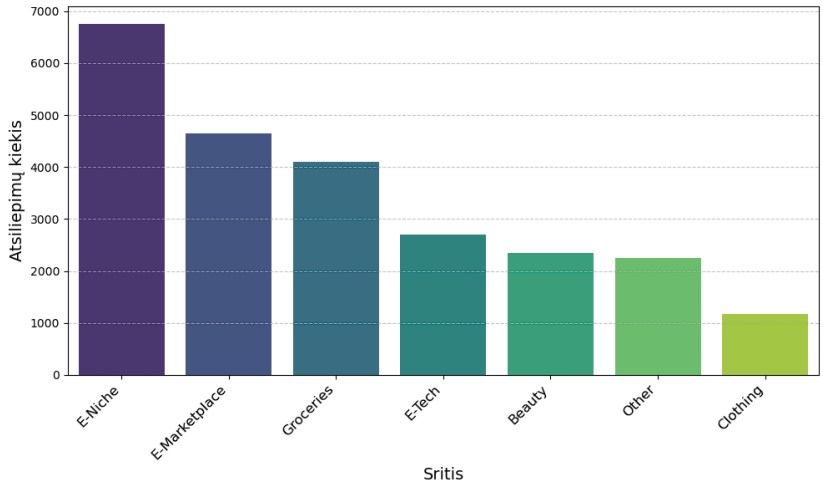
Duomenų rinkiniui papildyti bei atnaujinti bent 5000 tūkstančiais naujų atsiliepimų, pasitelktas įrankis „Apify“, nustatačius atsiliepimų paskelbimo laikotarpio rėžį nuo 2019-01-01 iki 2025-03-31 imtinai. Atsiliepimams surinkti pasirinkta „Google Maps“ verslų atsiliepimų platforma, kadangi minėtos „evertink.lt“ bei „Facebook“ verslo paskyros nebėra patikimas šaltinis surinkti didesniam kiekiui vartotojų atsiliepimų. Atsižvelgiant į tarpsritinės sentimentų analizės darbo tematiką, buvo išgaunami atsiliepimai iš skirtingų mažmeninės prekybos įmonių sričių, 5 didžiausiuose Lietuvos miestuose: didelės elektroninės bei fizinės parduotuvės („Pigu“, „Varlė“, „Senukai“), grožio prekės („Douglas“, „Drogas“, „Eurokos“, „Kristiana“), maisto prekės („Maxima“, „Lidl“, „IKI“) bei drabužiai („H&M“, „NewYorker“, „Reserved“, „Sportland“). Surinkus atsiliepimus, rankiniu būdu pašalinti nelietuviški atsiliepimai, bei duomenų faile pridėtas papildomas stulpelis, naudojamas identifikuoti mažmeninės prekybos sritis tolimesniuose eksperimentuose: elektroninės (bei fizinės) parduotuvės – (angl. *e-shops*); grožio prekės – (angl. *beauty*), maisto prekės – (angl. *groceries*) bei drabužiai – (angl. *clothing*). Po pirminio filtravimo, surinkti 5471 originaliam duomenų masyvui papildyti skirti internetiniai vartotojų atsiliepimai.

Prieš sujungiant duomenų rinkinius, originaliame duomenų masyve pašalintas klasių žymėjimo stulpelis bei pridėtas sritis identifikuojantis stulpelis. Rankiniu būdu peržiūrėjus įmones, esančias duomenų rinkinyje, nuspręsta praplėsti esamų sričių kategorijas, siekiant išvengti papildomo klasių disbalanso, kurį gali sukelti didelis skaičius atsiliepimų panašioms įmonėms, kurios patektų į tą pačią sritį. Sujungus duomenų rinkinius, bei atsižvelgiant į šį klasių disbalanso apribojimą, 3 lentelėje pateikiamos septynios, naujai sukurtos, galutiniuose eksperimentuose naudojamos mažmeninės prekybos vartotojų atsiliepimų sritys.

3 lentelė. Atnaujintame duomenų rinkinyje naudojamų atsiliepimų sričių apibendrinimas

Srities pavadinimas	**Srities apibūdinimas**	**Srities įmonių pavyzdžiai**	**Srities atsiliepimų skaičius**
E-Marketplace	Didelės elektroninės bei fizinės parduotuvės, siūlančios platų spektrą įvairių prekių.	„varle.lt“, „pigu.lt“, „1a.lt“	4657
E-Tech	Elektroninės bei fizinės parduotuvės, siūlančios kompiuterines bei elektronikos prekes.	„neriba.lt“, „elektromarkt.lt“, „gamecard.lt“	2693
E-Niche	Elektroninės bei fizinės parduotuvės, siūlančios specializuotas prekes bei reikmenis.	„Senukai“, „baldai1.lt“, „knygos.lt“	6757
Groceries	Elektroninės bei fizinės parduotuvės, siūlančios maisto produktus, maitinimo įstaigų atsiliepimai.	„Maxima“, „Lidl“, „Sushi Express“	4124
Clothing	Elektroninės bei fizinės parduotuvės, siūlančios drabužius.	„H&M“, „Reserved“, „New Yorker“	1174
Beauty	Elektroninės bei fizinės parduotuvės, siūlančios grožio prekes, parfumeriją.	„Douglas“, „Eurokos“, „KristiAna“	2353
Other	Įvairios paslaugos, prekybos centrai, degalinės, bankai, ir t. t.	„Circle K“, „Lietuvos zoologijos sodas“, „Danske Bank“	2251

Akcentuojama, kad duomenų rinkinio atsiliepimai buvo grupuoti, remiantis įmonės veiklos kontekstu bei siūlomų prekių pobūdžiu. Pavyzdžiui, „Senukai“ ir „knygos.lt“ priskirtos į *E-Niche* kategoriją , dėl specializuotos siūlomų prekių pasiūlos. Toliau, „Maxima“ ir „Sushi Express“ grupuotos *Groceries* srityje, dėl maisto prekių pasiūlos bei su jais susijusių panašių semantinių išraiškų vartotojų atsiliepimuose buvimo, kaip „*skanu*“, „*kokybiškas*“ ar „*didelis pasirinkimas*“. Atsižvelgiant į tai, kad šiame darbe orientuojamąsi į mažmeninę prekybą, įmonės, kurios pasižymėjo mišria prekių ir paslaugų struktūra, ir neatitiko darbo tematikos, buvo grupuojamos į *Other* sritį. Tokiu būdu išlaikomos aiškios eksperimento ribos bei koncentruojamąsi į mažmeninės prekybos specifiką, kadangi eksperimentuose *Other* sritis panaudojama išskirtinai modelių testavimui. Tokiu būdu galima įvertinti, kaip pasirinkti sentimentų analizės metodai geba generalizuoti į skirtingas bei mažai struktūriškai ir semantiškai panašias atsiliepimų sritis. Žemiau pateikiamas atskirų sukurtų sričių pasiskirstymas mažėjimo tvarka, pagal atsiliepimų skaičių (žr. 7 pav.).



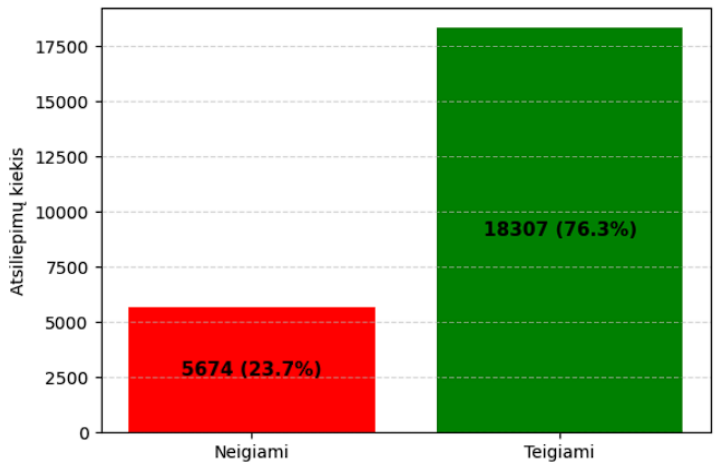
7 pav. Duomenų rinkinio vartotojų atsiliepimų skaičiaus pasiskirstymas, skirstyta pagal sritį

Siekiant paruošti turimą sujungtą duomenų masyvą binarinei sentimentų klasifikacijai, kitame žingsnyje atsiliepimams įvykdytas dviejų klasių žymų priskyrimas. Duomenų rinkinyje sukurtas naujas stulpelis, kuriame rankiniu būdu atsiliepimams priskirtos žymės: teigiamas, jeigu atsiliepimo įvertinimas siekė 4 arba 5 balus; neigiamas, jeigu atsiliepimo įvertinimas siekė 1, 2 arba 3 balus. Verta paminėti, jog žymint duomenis šiuo būdu atsiranda tam tikrų apribojimų. Atsiliepimuose išreiškiamas tekstas nebūtinai visada atitinka suteikiamą įvertinimą balais. Galimi atsiliepimai, kurie, nors ir turi priskirtą žemą įvertinimą balais (pavyzdžiui, 2), tačiau išreiškiamas pozityvus tonas atsiliepimo tekste. Kitu atveju, pastebimos neigiamos ypatybės tekste, nors suteiktas aukštas įvertinimas balais (pavyzdžiui, 5). Tokios klaidinančios atsiliepimų žymos gali kelti tam tikrą žymų triukšmą (angl. *label noise*), darantį minimalią įtaką modelių efektyvumui. Tačiau atsižvelgiama į tai, kad skirtinguose sentimentų klasifikacijos moksliniuose darbuose toks duomenų paruošimo metodas dažnai yra įprasta praktika, todėl šiame darbe pasirinkta išlaikyti rankiniu būdu priskirtas žymas, be papildomų metodų pritaikymo [50] [71].

Kitame žingsnyje atliktas duomenų paruošimas tolimesnei sentimentų analizei. Automatiniu būdu iš atsiliepimų tekstų pašalinami neinformatyvūs žodžiai (angl. *stopwords*), nepridedantys papildomos vertės modelių efektyvumui (pvz., „*anas*“, „*nagi*“, „*tegul*“ ir kt.). Visi esami simboliai paversti mažosiomis raidėmis, skyrybos ženklai atskirti tarpais, pašalinti skaitmenys, specialieji simboliai, jaustukai bei pašalinti papildomi tarpai tarp žodžių, siekiant sumažinti triukšmo kiekį duomenų rinkinyje. Taip pat pašalinami atsiliepimai su tuščiais teksto laukais. Po teksto valymo, naujame stulpelyje fiksuojamoms klasių žymoms priskirtos skaitinės reikšmės: neigiamoms priskirtos 0; teigiamoms priskirtos 1. Sukurtas naujas, atskiras stulpelis išvalytam tekstui. Siekiant užtikrinti lietuviškų simbolių palaikymą tolimesniems mašininio mokymosi etapams, išvalytas duomenų masyvas išsaugotas UTF-8 koduote. Galutiniame, tolimesnei tarpsritinei sentimentų analizei paruoštame duomenų rinkinyje – 23981 atsiliepimai.

3.2. Žvalgomoji analizė

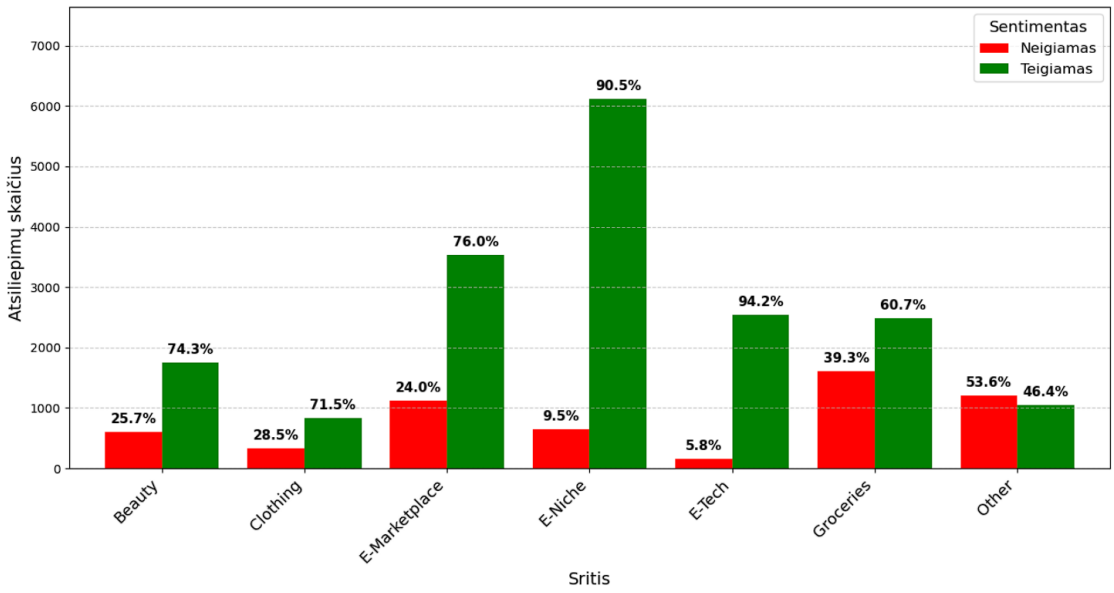
Šiame poskyryje atliekama žvalgomoji duomenų rinkinio analizė. Pateikiamas klasių balansas duomenų rinkinyje (žr. 8 pav.).



8 pav. Duomenų rinkinio klasių pasiskirstymas

Pastebima, jog duomenų rinkinyje dominuoja teigiama (1) klasė. Atsižvelgiant į šį faktą, priimtas sprendimas tyrimo rezultatuose naudoti atitinkamas vertinimo metrikas, tinkamas vertinti nesubalansuotus duomenų rinkinius.

Toliau pateikiamoje iliustracijoje (žr. 9 pav.) nurodomas sentimentų pasiskirstymas vartotojų atsiliepimuose atskirose srityse.



9 pav. Duomenų rinkinio klasių pasiskirstymas, skirstyta pagal sritį

Žvelgiant į atskirų sričių pasiskirstymą, didžiausias disbalansas pastebimas *E-Tech* bei *E-Niche* srityse, kuriose dominuoja teigiami sentimentai. Priešingai, *Other* srityje fiksuojamas didžiausias neigiamų sentimentų kiekis.

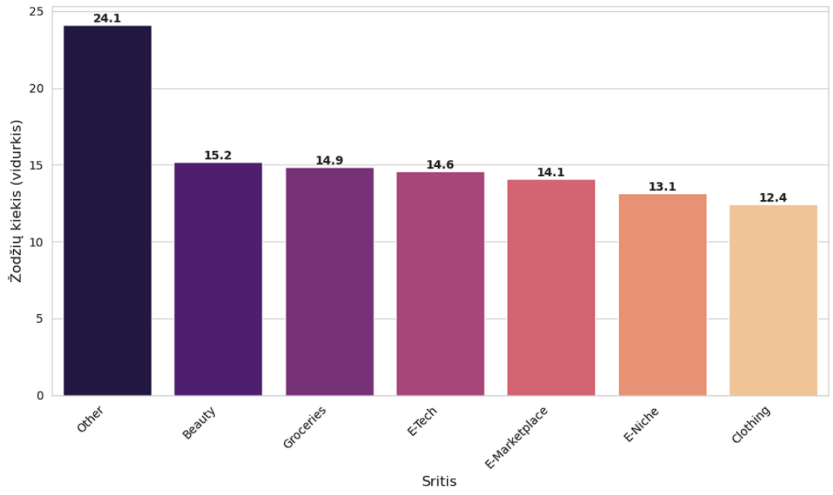
Toliau pateikiami žodžių debesys neigiamems atsiliepimams, pagal atskiras sritis. Šiuo metodu galima identifikuoti esmines klientų nepasitenkinimo tematikas (žr. 10 pav.).

10 pav. Žodžių debesų vizualizacijos sukurtos neigiamiems atsiliepimams, skirstyta pagal sritį

Pastebima, jog atskiroms sritims aktualūs skirtingi atitinkami esminiai žodžiai – pavyzdžiui, *Beauty* srityje figūruoja „*aptarnavimas*“, „*konsultantė*“, bei „*kvepalu*“. Kitoje srityje, *Groceries*, matoma, jog neigiamiuose atsiliepimuose dažnai pateikiami recenzuojamos įmonės pavadinimai bei

išsireiškimai apie kainas. *E-Tech* srityje minimos šios specializuotos srities prekės, t. y., kompiuterinė technika, kaip vienos iš nepasitenkinimo priežasčių. Iš 10 pav. taip pat matoma, jog kiekvienai iš sričių (išskyrus *Other*, kuri yra labiau orientuota į įvairias paslaugas, vietoves, bei pastebimi bendresni nepasitenkinimo identifikatoriai) išryškėja esminės problematinės sritys, kas leidžia įmonėms sudaryti atitinkamas strategines išvadas.

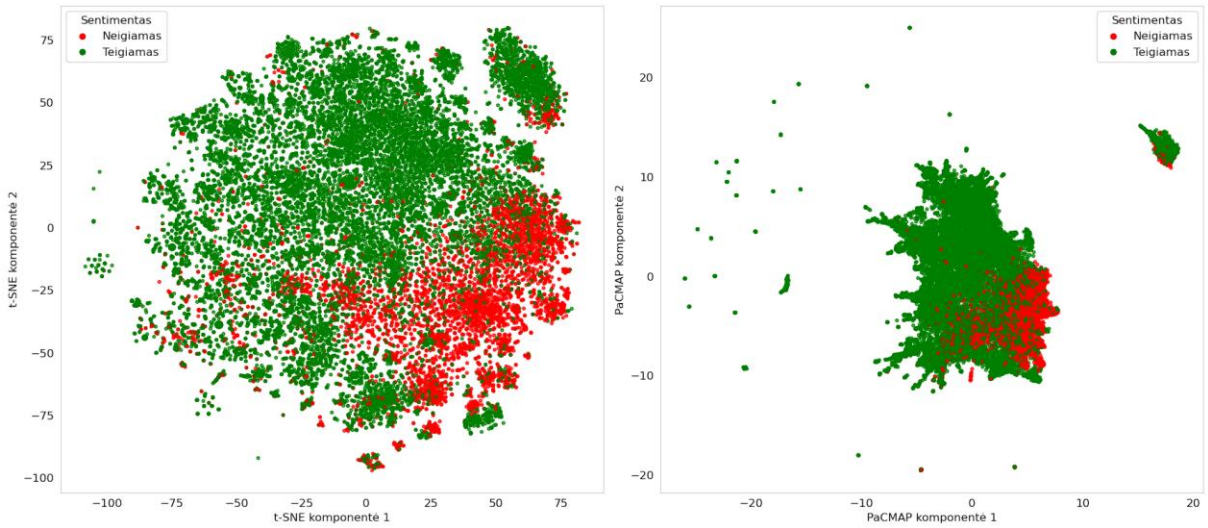
Siekiant įvairiapusiškai įvertinti sričių skirtumus, kitoje iliustracijoje (žr. 11 pav.) pateikiamas vidutinis atsiliepimo ilgis atskirose srityse.



11 pav. Duomenų rinkinio vidutinis atsiliepimo ilgis, skirstyta pagal sritį

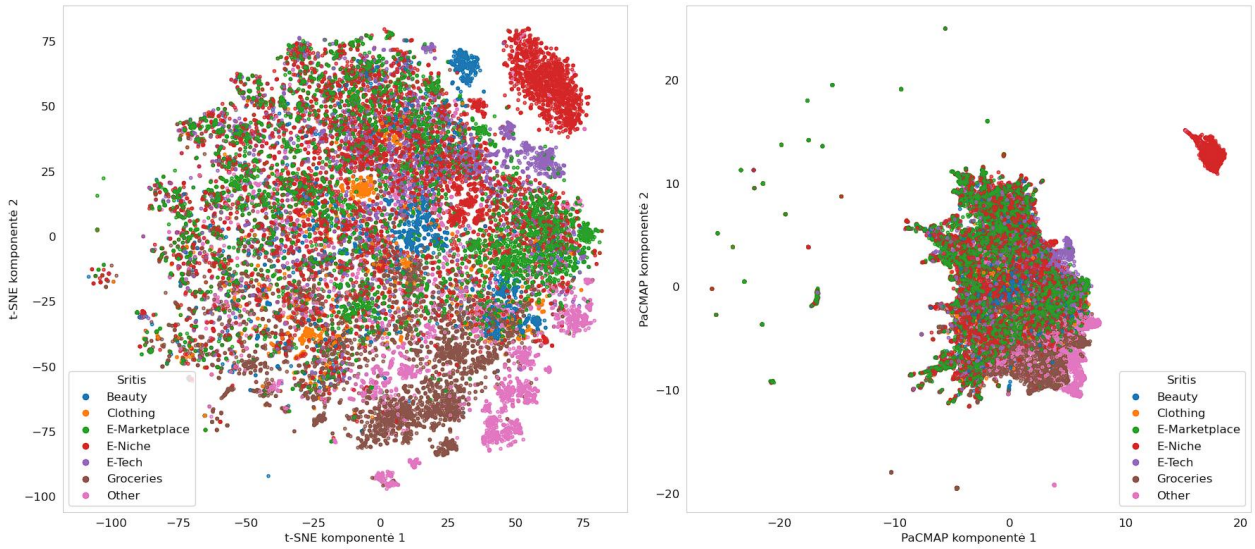
Didžiausias vidutinis žodžių kiekis atsiliepimuose fiksuojamas *Other* srityje. Tai gali būti siejama su šios srities įmonių specifika – t. y., įvairios paslaugų įmonės, bei kito pobūdžio institucijos, kaip bankai, degalinės ar prekybos centrai. Vartotojų patirtys šioje srityje dažnai yra sudėtingesnės, bei reikalaujančios detalesnio paaiškinimo. Šios srities atsiliepimai taip pat gali būti labiau pasakojamieji, o ne tiesioginės emocinės reakcijos, kas dažniau gali pasitaikyti perkant specializuotas prekes (pvz., elektroniką ar grožio prekes). Mažiausias žodžių kiekis – *Clothing* srityje, kur tikėtina, jog vartotojai apsiriboja trumpais, lakoniškais vertinimais („*prasta kokybė*“, „*netiko dydis*“), kurie perteikia emociją, tačiau nesuteikia daug papildomos kontekstinės informacijos.

Siekiant geriau suprasti duomenų semantinę struktūrą prieš atliekant tarpsritinius eksperimentus, atlikta vizualinė analizė, pasitelkiant du skirtingus dimensijų mažinimo metodus: *t-SNE* ir *PaCMAP*. Šioms vizualizacijoms generuoti panaudoti, tikėtina, efektyviausio vektorizavimo modelio (E5) sugeneruoti požymiai. Vizualizacijos (žr. 12, 13 pav.) leidžia preliminariai įvertinti, ar egzistuoja natūraliai susidarę klasteriai tarp skirtingų klasių ir sričių, bei identifikuoti galimą semantinę struktūrą tekstuose.



12 pav. *t-SNE* (kairėje) ir *PaCMAP* (dešinėje) vizualizacijos, skirstyta pagal sentimentus

Kairėje, *t-SNE* iliustracijoje, pastebima dalinė teigiamų (žalia) ir neigiamų (raudona) atsiliepimų atskirtis. Neigiami atsiliepimai dažniau pasiskirstę apatinėje, dešinėje dalyje. Tačiau klasės išlieka persidengusios – tai gali reikšti, kad skirtingų klasių atsiliepimai tam tikrais atvejais pasižymi panašia semantine struktūra. Tai būdinga vartotojų kuriamam turiniui, kur nuomonės ne visada yra vienareikšmiškos. Dešinėje, *PaCMAP* vizualizacijoje, sentimentų struktūra ryškesnė – teigiami ir neigiami atsiliepimai formuoja kompaktiškesnes grupes, kol teigiami atsiliepimai išsidėstę plačiau. Toliau pateikiamos t-SNE ir PaCMAP vizualizacijos, skirstant pagal atskiras sritis (žr. 13 pav.)



13 pav. *t-SNE* (kairėje) ir *PaCMAP* (dešinėje) vizualizacijos, skirstyta pagal sritis

Pastebima, jog tam tikros sritys (*Groceries*, *Other*, *E-Niche)* formuoja aiškesnes klasterines struktūras, bei kitos – *Clothing* ir *E-Marketplace*, yra pasiskirsčiusios plačiau. Tai gali rodyti, kad tam tikrų sričių tekstai pasižymi savitomis semantinėmis ypatybėmis, kurias vektorizacijos metodas geba išskirti. *PaCMAP* vizualizacija parodo kompaktiškesnius, tačiau labiau persidengiančius sričių pasiskirstymus nei *t-SNE*. Išsiskiria *E-Niche* sritis, kuri suformuoja izoliuotą, atskirą klasterį. Tai leidžia daryti prielaidą, kad šios srities atsiliepimai semantiškai reikšmingai skiriasi nuo kitų. Likusios sritys rodo didesnį semantinį panašumą, kas yra susiję su bendresniu atsiliepimų tonu, ar platesne siūlomų prekių ar paslaugų įvairove.

3.3. Tarpsritinės sentimentų analizės rezultatai

Tarpsritinės sentimentų analizės rezultatai buvo suskirstyti pagal dimensionalumą į dvi atskiras eksperimentines grupes – 768D ir 1024D. Visos LSA vektorizavimo ir klasifikatorių kombinacijos, bei *XLM-RoBERTa* (*E2E*) klasifikatorius buvo įtraukti į abi eksperimentines grupes dėl šių metodų universalumo bei lankstumo. *GTE* (768D), *XLM-RoBERTa* (su generuotais įterpiniais) (768D), *Jina* (1024D) ir *E5* (1024D) metodai naudojami atitinkamose grupėse. Tokia struktūra leidžiama palyginti metodus, atsižvelgiant į jų taikymo pobūdį bei galimybes realiose analizės situacijose.

3.3.1. 768D duomenų rinkinys

Šiame poskyryje pateikiamos 768D eksperimentinės grupės rezultatai. Siekiant užtikrinti nuoseklumą bei optimalius rezultatus, prieš vykdant tarpsritinius eksperimentus, kiekvienai vektorizavimo ir klasifikatoriaus kombinacijai atliktas parametrų priderinimas naudojant tinklelio paiešką. Papildomai, *XLM-RoBERTa* (su generuotais įterpiniais) ir reguliarizuotos logistinės regresijos metodui pridėtas minimalus ($1 * 10^{-16}$) mašininio tikslumo triukšmas (angl. *machine precision noise*), siekiant išvengti konvergavimo problemų.

Užfiksuoti kiekvienos kombinacijos tarpsritinių eksperimentų apskaičiavimo laikai. Trumpiausias skaičiavimo laikas pasiektas su *LSA* ir reguliarizuota logistine regresija – 23 sekundės. Ilgiausias apskaičiavimo laikas – *XLM-RoBERTa* (su generuotais įterpiniais) bei reguliarizuota logistine regresija – 3669 sekundės, t. y., 61 minutė.

Toliau pateikiama aukščiausių užfiksuotų vektorizacijos ir klasifikavimo modelių rezultatai šiame eksperimente (žr. 4 lentelė). Siekiant užtikrinti optimalų rezultatų vertinimą, pasitelkiami PRC AUC ir ROC AUC įverčiai, atsižvelgiant į nesubalansuotą duomenų rinkinį.

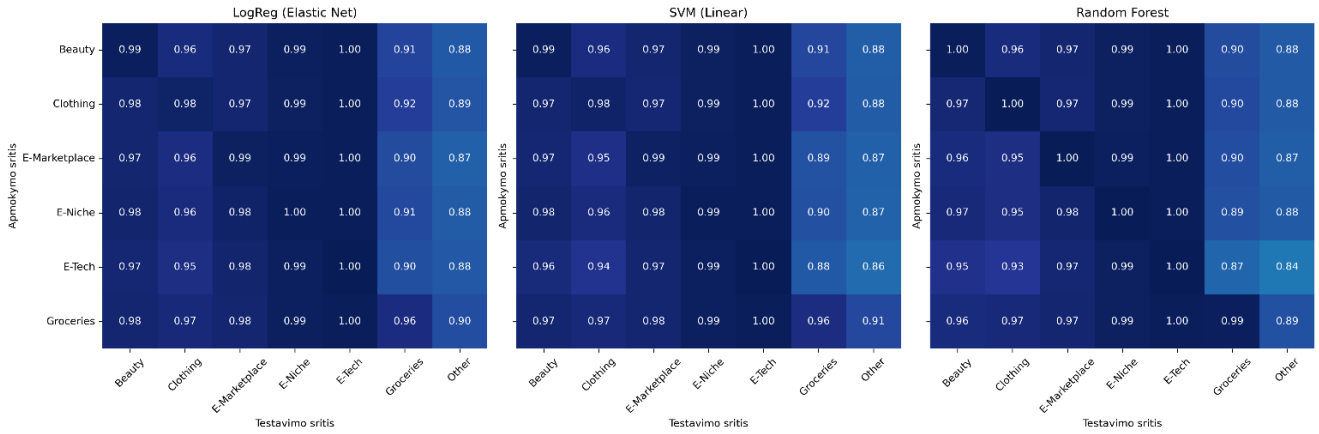
4 lentelė. Geriausi vektorizavimo ir klasifikavimo modelių rezultatai 768D eksperimentinėje grupėje (paryškinti geriausi klasifikavimo algoritmai kiekvienam vektorizavimo metodui)

	LogReg (*Elastic Net*)		***SVM Linear***		***Random Forest***		***XLM-R* (*End-to-End*)**	
	PRC AUC	ROC AUC	PRC AUC	ROC AUC	PRC AUC	ROC AUC	PRC AUC	ROC AUC
GTE	**0,9977**	**0,9643**	0,9975	0,9614	0,9970	0,9560	-	-
LSA	**0,9977**	**0,9645**	0,9967	0,9513	0,9963	0,9461	-	-
XLM-R	0,9966	0,9472	**0,9983**	**0,9745**	0,9962	0,9438	0,9981	0,9670

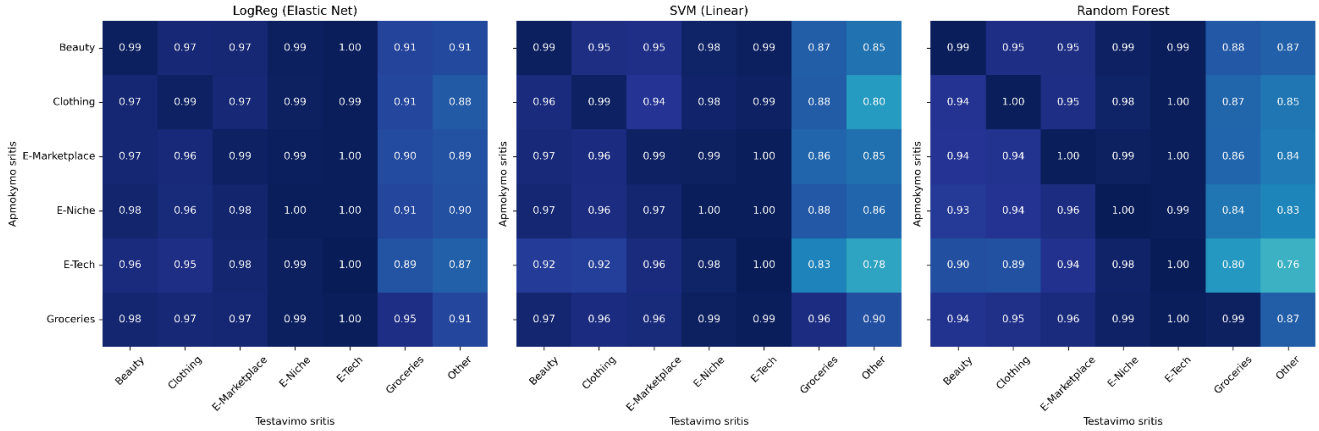
Atskirai, prie *XLM-RoBERTa* vektorizavimo metodo pateikti *XLM-RoBERTa* (*E2E*) klasifikatoriaus įverčiai, kadangi šiam modeliui nėra naudojami atskiri klasifikatoriai. Iš pateiktų rezultatų matoma, jog 768D duomenų rinkinyje geriausias rezultatas pasiektas naudojant *XLM-RoBERTa* (su generuotais įterpiniais) ir tiesinį *SVM* klasifikatorių. Didžiausi PRC AUC ir ROC AUC įverčiai *GTE* ir *LSA* vektorizavimo metodams pasiekiami naudojant reguliarizuotos logistinės regresijos klasifikatorių. Atskirai vertinamas *XLM-RoBERTa* (*E2E*) klasifikatorius taip pat pasiekia itin konkurencingus rezultatus. Nors *Random Forest* klasifikatoriumi nepasiekiami aukščiausi įverčiai nei su vienu vektorizavimo metodu, tačiau rezultatai išlieka pakankamai aukšti, ypač PRC AUC reikšmėms.

Toliau vertinamas vektorizacijos ir klasifikavimo kombinacijų, bei atskiro klasifikatoriaus, *XLM-RoBERTA* (*E2E*) efektyvumas tarpsritinės sentimentų analizės scenarijuose. Analizuojama metodikų

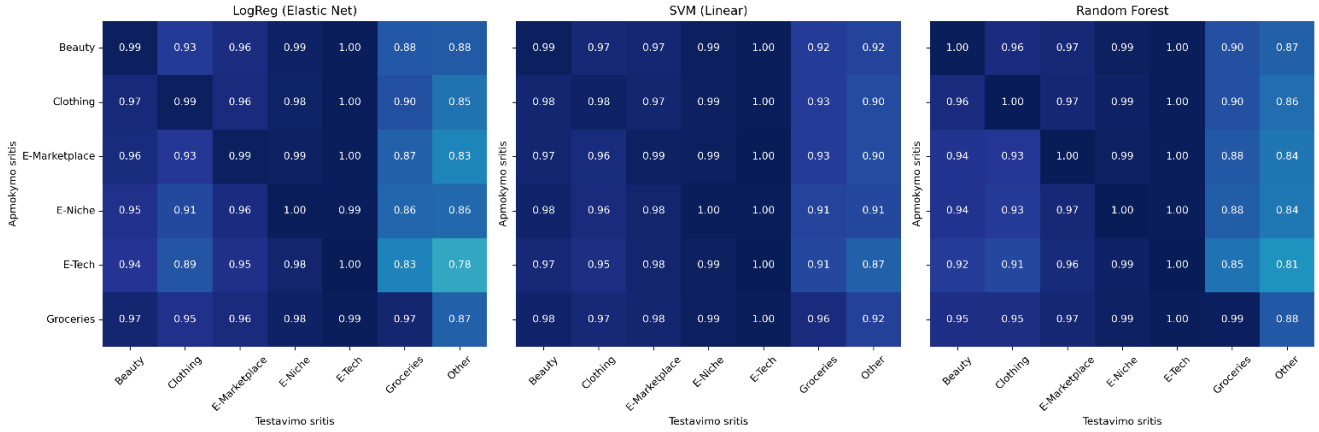
generalizacija, pateikiant šilumos žemėlapių (angl. *heatmap*) vizualizacijas atskiriems vektorizacijų ir klasifikatorių metodams (žr. 14, 15, 16, 17 pav.)



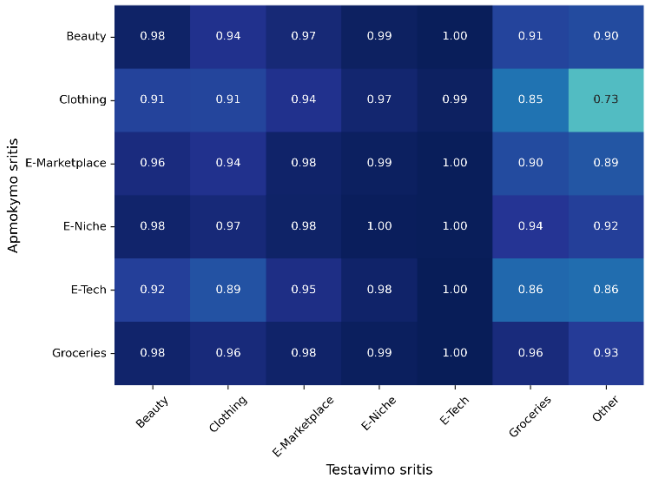
14 pav. *GTE* vektorizacijos PRC AUC šiluminiai žemėlapiai, skirstyti pagal klasifikatorių (geriausi rezultatai – su reguliarizuota logistine regresija)



15 pav. *LSA* vektorizacijos PRC AUC šiluminiai žemėlapiai, skirstyti pagal klasifikatorių (geriausi rezultatai – su reguliarizuota logistine regresija)



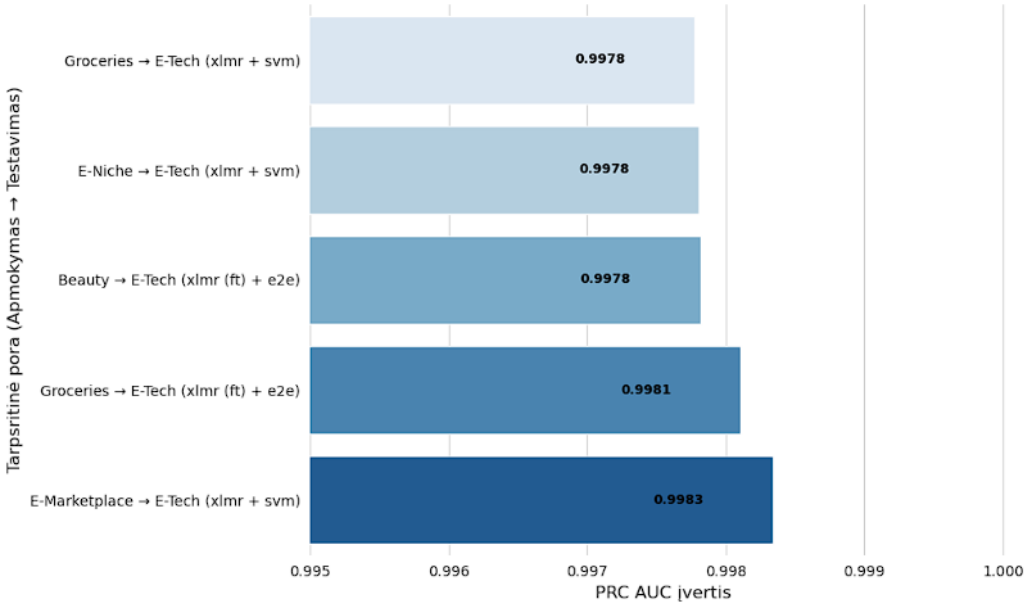
16 pav. *XLM-RoBERTa* (su generuotais įterpiniais) vektorizacijos PRC AUC šiluminiai žemėlapiai, skirstyti pagal klasifikatorių (geriausi rezultatai – su tiesiniu *SVM*)



17 Pav. *XLM-RoBERTa* (*End-to-End*) klasifikatoriaus PRC AUC šiluminis žemėlapis

Šiluminių žemėlapių vizualizacijos parodo, kad visose 768D eksperimentinės grupės kombinacijose pasiekiami itin aukšti PRC AUC įverčiai, daugeliu atvejų viršijant 0,90. Vienodose srityse apmokomi ir testuojami metodai išsiskiria beveik maksimalių reikšmių rezultatais (0,98–1,00), kol tarpsritiniai eksperimentai taip pat išlaiko aukštą stabilumą. Pastebima tendencija, jog *Groceries*, kaip apmokymo srities rezultatai, testuojant ant kitų sričių yra aukštesni, lyginant su kitomis apmokymo sritimis, per visus eksperimento metodus. Kaip ir tikėtasi, *Other* yra viena sudėtingiausių testavimo sričių, siekianti 0,73 PRC AUC (*Clothing-Other)*, kaip bendra mažiausia reikšmė tarp visų metodikų. Tokie rezultatai leidžia daryti prielaidą, jog 768D duomenų rinkinys ir naudoti metodai gali būti efektyviai pritaikomi tarpsritiniuose sentimentų analizės scenarijuose mažmeninės prekybos srityje.

Toliau pateikiamos 5 tarpsritinės sentimentų analizės apmokymo bei testavimo poros su didžiausiais PRC AUC įverčiais (žr. 18 pav.).

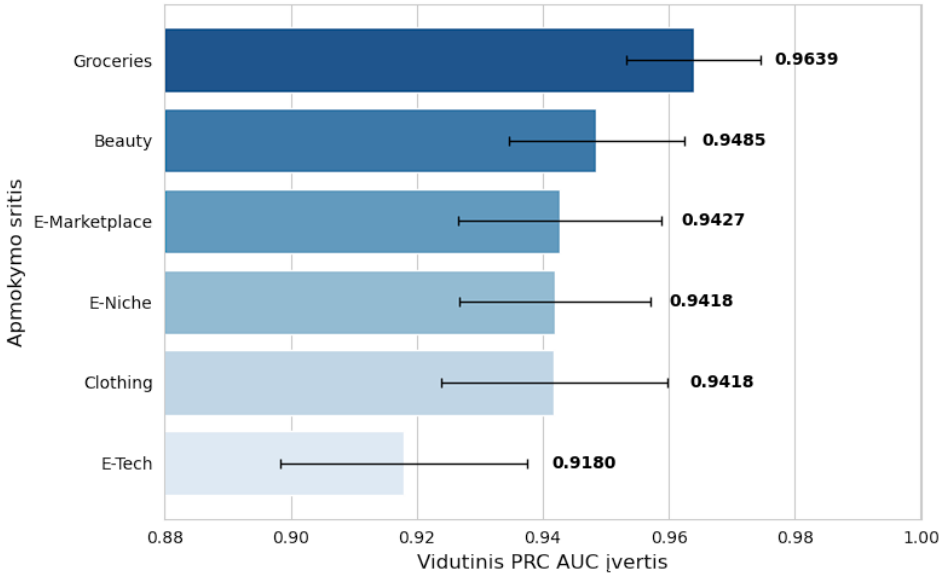


18 pav. Geriausios 5 poros 768D eksperimentinėje grupėje, pagal aukščiausias PRC AUC reikšmes

Pastebima, kad 768D duomenų rinkiniui, aukščiausias PRC AUC rezultatas (0,9983) fiksuojamas naudojant *XLM-RoBERTa* (su generuotais įterpiniais) kartu su tiesiniu SVM klasifikatoriumi. Tarp 5 geriausių porų, dominuoja *E-Tech* testavimo sritis. Matoma, jog efektyviausi rezultatai pasiekiami su

XLM-RoBERTa (tiek generuotais įterpiniais, tiek *E2E* metodu). Toks faktas pabrėžia šio daugiakalbio metodo efektyvumą, pritaikant metodą tarpsritinės sentimentų analizės scenarijams, net ir mažai resursų turinčiose kalbose, kaip lietuvių.

Siekiant identifikuoti efektyviausią tarpsritinės sentimentų analizės porą šiam duomenų rinkiniui, pateikiamos apmokymo sritys, pagal vidutines PRC AUC reikšmės (žr. 19 pav.).



19 pav. Geriausios apmokymo sritys 768D eksperimentinėje grupėje, pagal vidutines PRC AUC reikšmes

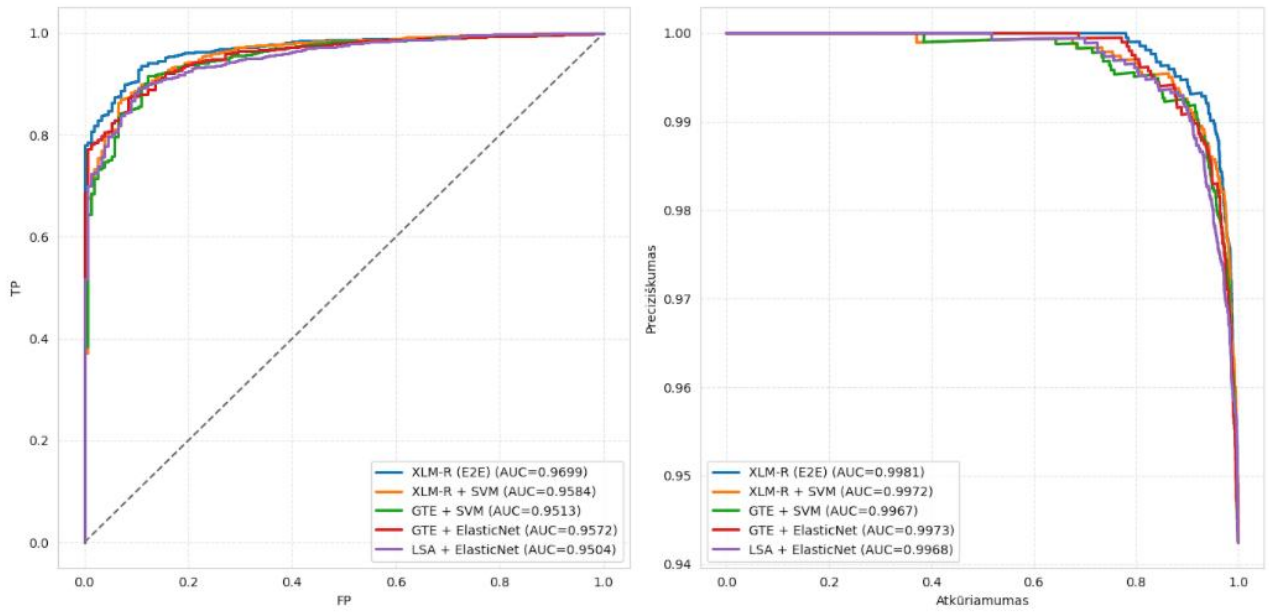
Matoma, jog *Groceries* yra efektyviausia apmokymo sritis pagal vidutinę PRC AUC reikšmę (0,9639±0,0107), kas buvo pastebima šiluminių žemėlapių vizualizacijose. Tai gali būti paaiškinama dėl plataus spektro semantinių išsireiškimų, naudojamų šios srities atsiliepimams, dėl to analizė gali būti efektyviau pritaikoma per skirtingų sričių kontekstus.

Toliau, siekiant atrasti efektyviausią tarpsritinės sentimentų analizės porą 768D grupėje, (žr. 5 lentelė) pateikiamos 5 poros, su nustatyta geriausia apmokymo sritimi – *Groceries*. Pridedamas Kappa koeficiento rodiklis, siekiant įvertinti, kaip modelis geba generalizuoti tarp skirtingų sričių, neapsiribojant dominuojančios klasės dažniu.

5 lentelė. Geriausios poros 768D eksperimentinėje grupėje, su *Groceries* apmokymo sritimi

Apmokymo-testavimo pora (metodas)	**PRC AUC**	**ROC AUC**	**Kappa koeficientas**
Groceries - E-Tech (*XLM-RoBERTa* (*E2E*))	0,9981	0,9699	0,5054
Groceries - E-Tech (*XLM-RoBERTa + SVM*)	0,9972	0,9684	0,3815
Groceries - E-Tech (*GTE + SVM*)	0,9967	0,9513	0,3680
Groceries - E-Tech (*GTE + LogReg* (*Elastic Net*))	0,9973	0,9572	0,3366
Groceries - E-Tech (*LSA + LogReg* (*Elastic Net*))	0,9968	0,9504	0,3548

Penkiose geriausiose porose testavimo sritis visais atvejais yra *E-Tech*. Tai gali būti paaiškinama šios testavimo srities atsiliepimų semantiniu panašumu į apmokymo sritį, arba mažai išsiskiriančiais, labiau bendriniais vartotojų išsireiškimais testavimo srityje. Nustačius geriausias poras 768D eksperimentinėje grupėje, tolimesniam palyginimui, poros vizualizuojamos su ROC ir PR kreivėmis (žr. 20 pav.).



20 pav. ROC (kairėje) ir PR (dešinėje) kreivės geriausioms poroms 768D eksperimentinėje grupėje, *Groceries* apmokymo sritis, *E-Tech* testavimo sritis

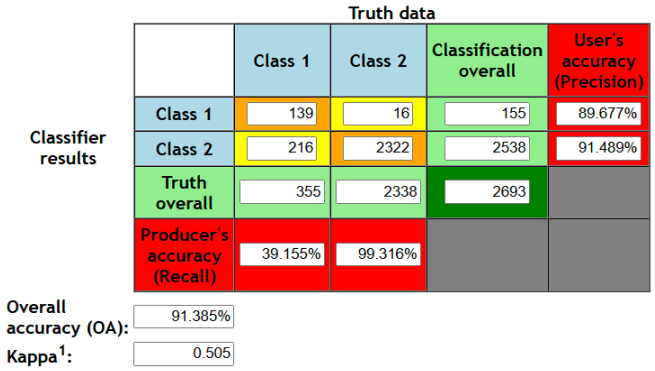
ROC kreivėje pastebima, kad *XLM-R* (*E2E*) klasifikatorius yra arčiausiai kairiojo viršutinio taško, todėl šis metodas yra užfiksuojamas kaip efektyviausias 768D duomenų rinkinyje. *XLM-RoBERTa* (*E2E*) klasifikatoriaus efektyvumas šiame duomenų rinkinyje taip patvirtinamas su PR kreive (21 pav.) Šiam metodui, lyginant su kitais pavaizduotais, pastebimas mažiausias preciziškumo sumažėjimas, didėjant atkūriamumo reikšmėms. Iš pateiktų ROC ir PR kreivių grafikų patvirtinama, kad tiksliausias bei efektyviausias klasifikavimo metodas 768D duomenų rinkiniui yra *XLM-RoBERTa* (*E2E*), pritaikytas *Groceries - E-Tech* poroje. Papildomai (žr. 6 lentelė) apžvelgiamos likusios poros su geriausiu 768D eksperimentinės grupės metodu, siekiant įvertinti metodikos gebėjimus, testuojant kitose srityse.

6 lentelė. Likusios poros efektyviausiam metodui (*XLM-RoBERTa* (*E2E*)) ir geriausiai apmokymo sričiai, *Groceries*, 768D eksperimentinėje grupėje

Apmokymo-testavimo pora	**PRC AUC**	**ROC AUC**
Groceries-E-Niche	0,9924	0,9699
Groceries-E-Marketplace	0,9808	0,9385
Groceries-Beauty	0,9796	0,9565
Groceries-Clothing	0,9639	0,9407
Groceries-Other	0,9261	0,9295

Pagal 6 lentelėje pateiktus geriausio šios eksperimentinės grupės metodo rezultatus, pastebimi aukšti testavimo rezultatai kitoms duomenų rinkinio sritims, apmokant *XLM-RoBERTa (E2E)* su *Groceries* sritimi. Aukštos PRC AUC ir ROC AUC reikšmės, testuojant tokias sritis kaip *E-Niche* ir *E-Marketplace*, potencialiai gali būti siejamos su apmokymo ir testavimo srityse dominuojančiais bendrais teminiais bruožais arba stilistiniais vartotojų išreiškiamų nuomonių panašumais. Tačiau šiuo metodu taip pat fiksuojami aukšti rezultatai testuojant mažiau semantiškai susijusias sritis, tokias kaip *Beauty* ir *Clothing*. Verta pabrėžti, kad *XLM-RoBERTa* (*E2E*) modelis geba efektyviai generalizuoti net ir į nesusijusias sritis, kaip *Other*, kur taip pat pasiekiami patenkinami rezultatai.

Toliau pateikiama efektyviausio 768D rinkinyje identifikuoto metodo, *XLM-RoBERTa* (*E2E*) bei geriausios tarpsritinės analizės poros, *Groceries - E-Tech* sumaišymo matrica. Siekiama identifikuoti, kaip geriausia šios eksperimentinės grupės kombinacija prognozuoja klasių žymas, lyginant geriausio metodo rezultatus prognozes bei tikrąsias duomenų rinkinio žymas (žr. 21 pav.).



21 pav. 768D eksperimentinės grupės geriausios kombinacijos sumaišymo matrica

Sumaišymo matricoje pastebima, kad šios eksperimentinės grupės geriausiu metodu ir pora pasiekiamas 91,385 % bendras tikslumas bei vidutinis Kappa koeficiento sutapimo lygis (0,505). Naudojant *XLM-RoBERTa* (*E2E*) šioje tarpsritinėje poroje, pasiekiami aukšti rezultatai identifikuojant teigiamą (1) klasę, su 99,316 % preciziškumu ir 91,489 % atkūriamumu. Tačiau matoma, kad metodo efektyvumas yra mažesnis, identifikuojant neigiamas (0) klases, kur atkūriamumo rodiklis siekia tik 39,155 %. Neigiamos klasės preciziškumas (89,677 %) išlieka sąlyginai aukštas – tai rodo, kad identifikuotos neigiamos klasės šios grupės geriausiu metodu dažniausiai būna parenkamos korektiškai.

3.3.2. 1024D duomenų rinkinys

Šiame poskyryje pateikiamos 1024D eksperimentinės grupės rezultatai. Norint užtikrinti nuoseklumą, prieš vykdant tarpsritinius eksperimentus, kiekvienai vektorizavimo ir klasifikatoriaus kombinacijai atliktas parametrų priderinimas naudojant tinklelio paiešką (angl. *grid search*).

Taip pat užfiksuoti kiekvienos kombinacijos tarpsritinių eksperimentų apskaičiavimo laikai. Trumpiausias skaičiavimo laikas pasiektas su *LSA* ir reguliarizuota logistine regresija – 23 sekundės. Ilgiausias apskaičiavimo laikas užfiksuotas apmokant *XLM-RoBERTa* (*E2E*) – 552 sekundės, t. y., 9,2 minutės.

Pateikiami aukščiausi fiksuoti vektorizacijos ir klasifikavimo modelių rezultatai šioje eksperimentinėje grupėje (žr. 7 lentelė). Vertinimui pasitelkiami PRC AUC ir ROC AUC įverčiai, atsižvelgiant į nesubalansuotą duomenų rinkinį.

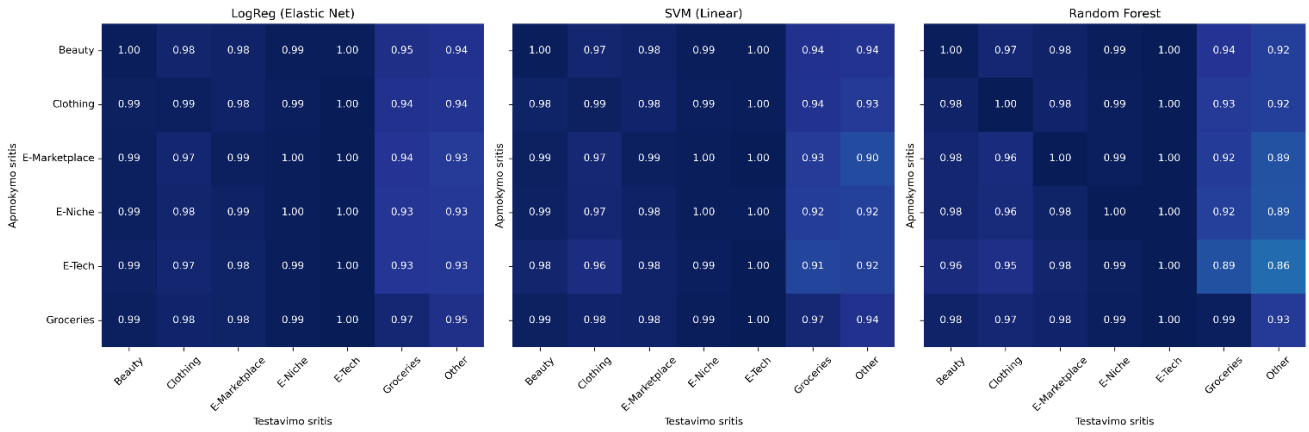
7 lentelė. Geriausi vektorizavimo ir klasifikavimo modelių rezultatai 1024D eksperimentinėje grupėje (paryškinti geriausi klasifikavimo algoritmai kiekvienam vektorizavimo metodui)

	LogReg (*Elastic Net*)		***SVM Linear***		***Random Forest***		***XLM-R* (*End-to-End*)**	
	PRC AUC	ROC AUC	PRC AUC	ROC AUC	PRC AUC	ROC AUC	PRC AUC	ROC AUC
Jina	**0,9987**	**0,9799**	0,9987	0,9795	0,9985	0,9777	-	-
E5	**0,9987**	**0,9804**	0,9984	0,9752	0,9986	0,9788	-	-

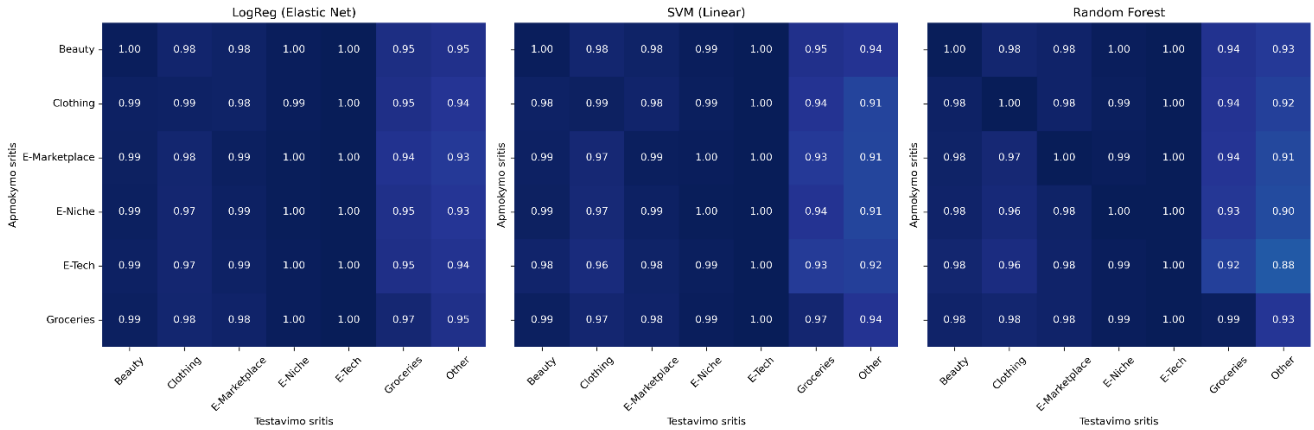
	LogReg (*Elastic Net*)		***SVM Linear***		***Random Forest***		***XLM-R* (*End-to-End*)**	
LSA	**0,9977**	**0,9645**	0,9967	0,9513	0,9963	0,9461	-	-
XLM-R	-	-	-	-	-	-	0,9981	0,9670

Atskirame stulpelyje pateikti *XLM-RoBERTa* (*E2E*) klasifikatoriaus įverčiai, kadangi šiam modeliui nėra naudojami atskiri klasifikatoriai. Geriausiuose šios grupės rezultatuose pastebima, kad 1024D duomenų rinkinyje aukščiausi įverčiai pasiekti naudojant *E5* ir reguliarizuotos logistinės regresijos klasifikatorių. Likusiems *Jina* ir *LSA* vektorizavimo metodams didžiausi PRC AUC ir ROC AUC pasiekiami taip pat naudojant reguliarizuotos logistinės regresijos klasifikatorių. Nors *SVM* ir *Random Forest* klasifikatoriais nepasiekiami aukščiausi įverčiai nei su vienu vektorizavimo metodu, tačiau rezultatai išlieka pakankamai aukšti, ypač PRC AUC reikšmėms.

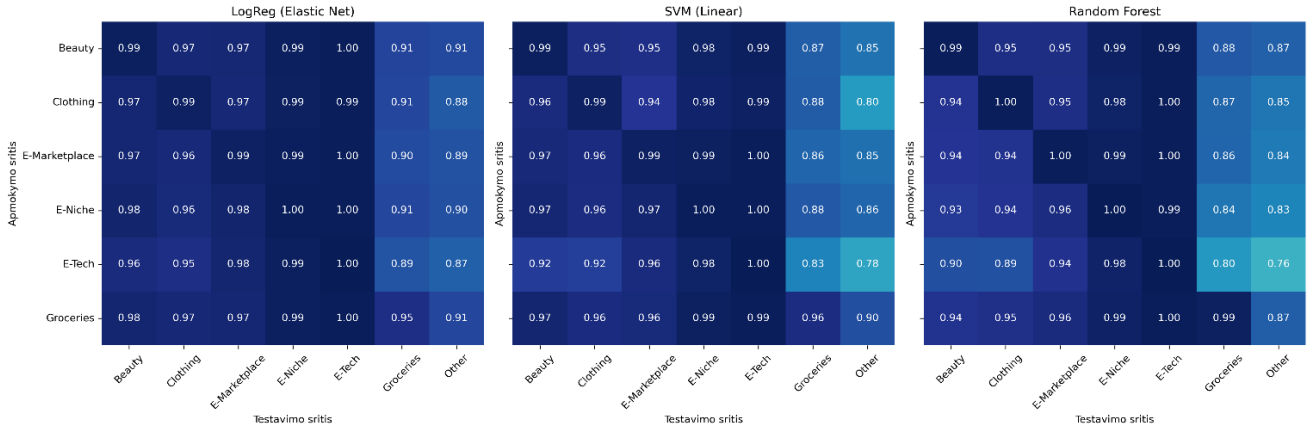
Toliau vertinamas atskirų metodų efektyvumas tarpsritinės sentimentų analizės scenarijuose. Analizuojama metodikų generalizacija, pateikiant šiluminių žemėlapių vizualizacijas atskiriems metodams (žr. 22, 23, 24, 25 pav.).



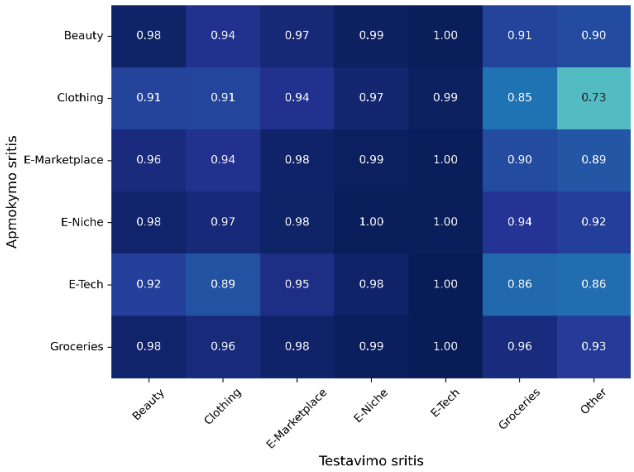
22 pav. *Jina* vektorizacijos PRC AUC šiluminiai žemėlapiai, skirstyti pagal klasifikatorių (geriausi rezultatai – su reguliarizuota logistine regresija)



23 pav. *E5* vektorizacijos PRC AUC šiluminiai žemėlapiai, skirstyti pagal klasifikatorių (geriausi rezultatai – su reguliarizuota logistine regresija)



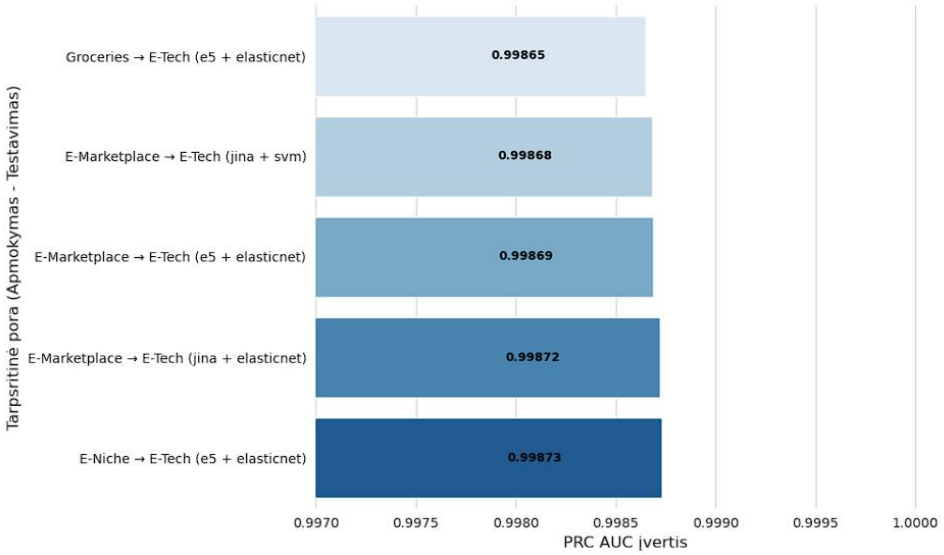
24 pav. *LSA* vektorizacijos PRC AUC šiluminiai žemėlapiai, skirstyti pagal klasifikatorių (geriausi rezultatai – su reguliarizuota logistine regresija)



25 pav. *XLM-RoBERTa* (*End-to-End*) klasifikatoriaus PRC AUC šiluminis žemėlapis

Pateiktose šiluminių žemėlapių vizualizacijose matoma, kad visose 1024D eksperimentinės grupės kombinacijose pasiekiami aukšti PRC AUC įverčiai (daugeliu atvejų viršijant 0,95). Beveik visais atvejais, vienodose srityse apmokomi ir testuojami metodai išsiskiria maksimalių reikšmių rezultatais (0,98–1,00). Tarpsritiniai eksperimentai išlaiko itin aukštą stabilumą – kas ypač pastebima su *Jina* ir *E5* metodais, kuriuose prasčiausia tarpsritinė pora (*E-Tech – Other*, su *Jina* ir *Random Forest* klasifikatoriumi) siekia 0,86 PRC AUC. Analogiškai, kaip ir 768D eksperimentinėje grupėje, čia taip pat pastebima tendencija, jog *Groceries*, kaip apmokymo srities rezultatai, testuojant ant kitų sričių yra aukštesni, lyginant su kitomis apmokymo sritimis, per visus eksperimento metodus. Kaip ir tikėtasi, *Other* yra viena sudėtingiausių apmokymo sričių, siekianti 0,73 (*Clothing - Other)*, kaip bendra mažiausia reikšmė tarp visų metodikų. Tarpsritinių porų rezultatai leidžia daryti prielaidą, jog 1024D duomenų rinkinys ir naudoti metodai gali būti dar efektyviau pritaikomi tarpsritiniuose sentimentų analizės scenarijuose, mažmeninės prekybos srityje.

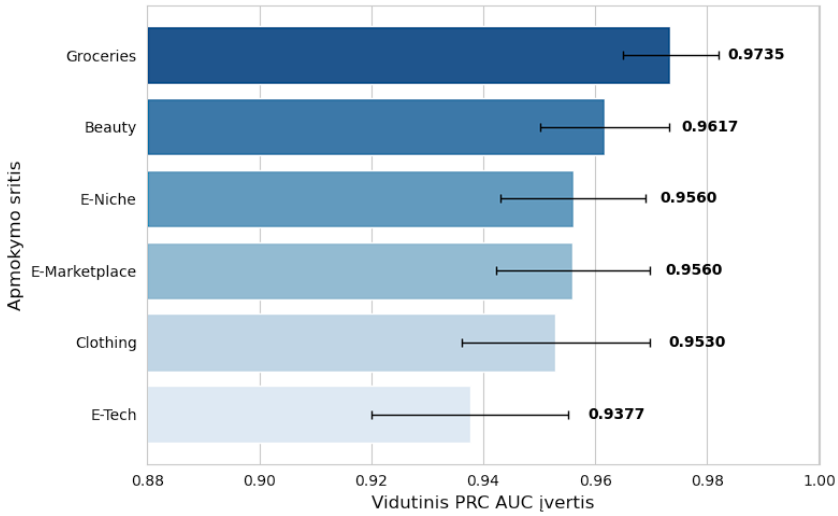
Toliau pateikiamos 5 tarpsritinės sentimentų analizės apmokymo bei testavimo poros bei jų metodai su didžiausiais PRC AUC įverčiais, (26 pav.).



26 pav. Geriausios 5 poros 1024D eksperimentinėje grupėje, pagal PRC AUC reikšmes

1024D eksperimentinėje grupėje aukščiausias PRC AUC rezultatas (0,99873) fiksuojamas naudojant *E5* vektorizavimo modelį kartu su reguliarizuota logistine regresija. Geriausiose porose, analogiškai, kaip ir 768D eksperimentinėje grupėje, dominuoja *E-Tech* testavimo sritis. Pastebima, kad efektyviausi rezultatai pasiekiami su *Jina* ir *E5* vektorizavimo metodais, kas pabrėžia šių modernių žodžių įterpinių generavimo modelių efektyvumą.

Toliau, siekiamia identifikuoti efektyviausią tarpsritinės sentimentų analizės porą 1024D eksperimentinei grupei – pateikiamos apmokymo sritys, rikiuojant pagal vidutines PRC AUC reikšmės (žr. 27 pav.).



27 pav. Geriausios apmokymo sritys 1024D eksperimentinėje grupėje, pagal vidutines PRC AUC reikšmes

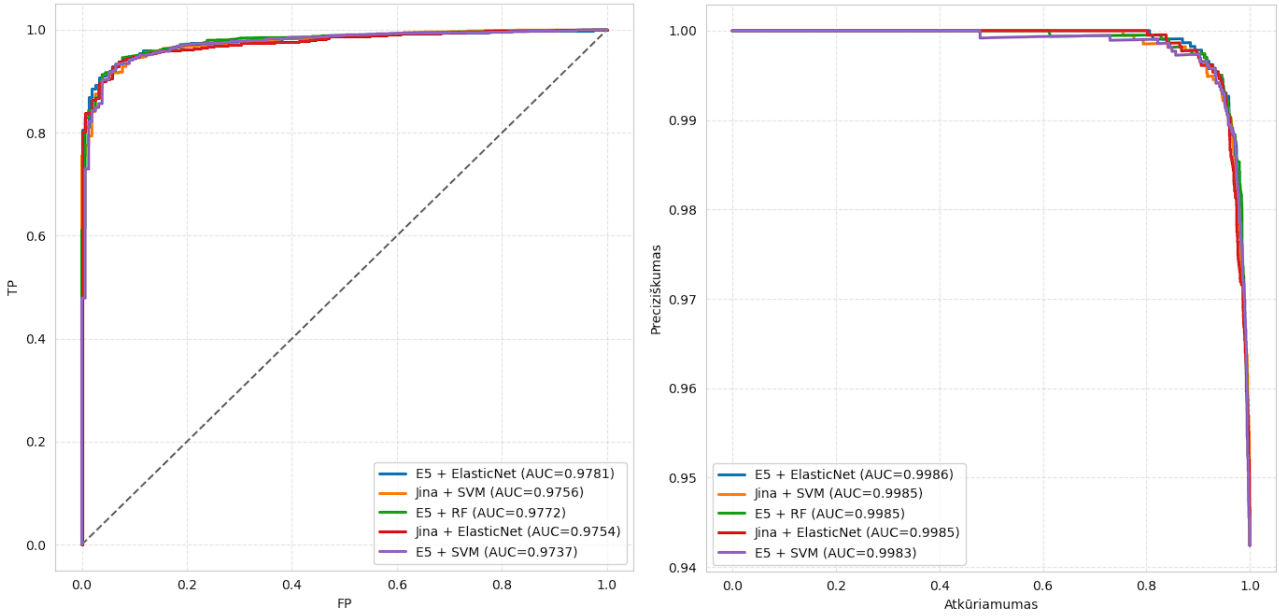
Pagal 27 pav. pastebima, jog *Groceries* yra efektyviausia apmokymo sritis (0,9735±0,0085). Kaip ir 768D eksperimentinėje grupėje, tai gali būti paaiškinama dėl platesnio semantinių išsireiškimų spektro ir bendresnio naudojimo terminų, naudojamų būtent šios srities atsiliepimams, todėl analizė su šia apmokymo sritimi gali būti efektyviau pritaikoma per skirtingų sričių kontekstus.

Siekiant atrasti efektyviausią tarpsritinės sentimentų analizės porą 1024D rinkinyje, (žr. 8 lentelė) pateikiamos 5 poros, su nustatyta geriausia apmokymo sritimi – *Groceries*.

8 lentelė. Geriausios poros 1024D eksperimentinėje grupėje, su *Groceries* apmokymo sritimi

Apmokymo-testavimo pora (metodas)	**PRC AUC**	**ROC AUC**	**Kappa koeficientas**
Groceries - E-Tech (*E5 + LogReg* (*Elastic Net*))	0,9986	0,9780	0,5587
Groceries - E-Tech (*Jina + SVM*)	0,9985	0,9770	0,4983
Groceries - E-Tech (*E5 + RF*)	0,9985	0,9772	0,5246
Groceries - E-Tech (*Jina + LogReg* (*Elastic Net*))	0,9984	0,9753	0,5064
Groceries - E-Tech (*E5 + SVM*)	0,9984	0,9752	0,5616

Nustačius geriausias poras 1024D rinkinyje, tolimesniam palyginimui, jos vizualizuojamos su ROC ir PR kreivėmis (žr. 28 pav.).



28 pav. ROC (kairėje) ir PR (dešinėje) kreivės geriausioms poroms 1024D eksperimentinėje grupėje, *Groceries* apmokymo sritis, *E-Tech* testavimo sritis

ROC kreivėje pastebima, kad *E5* ir reguliarizuotos logistinės regresijos metodas yra arčiausiai kairiojo viršutinio taško, todėl šis metodas yra užfiksuojamas kaip efektyviausias 1024D duomenų rinkinyje. PR kreivėje *E5* ir reguliarizuotos logistinės regresijos metodui taip pat pastebimas mažiausias preciziškumo sumažėjimas, didėjant atkūriamumo reikšmėms. Bendrai, ROC ir PR kreivėse pastebima itin nedidelė atskirtis tarp metodų, kas leidžia daryti išvadą, kad šiuo dimensionalumu vertinti metodai pasiekia aukštesnius rezultatus, nei 768D duomenų rinkinyje išbandyti metodai. Iš pateiktų ROC ir PR kreivių grafikų patvirtinama, kad geriausias klasifikavimo metodas 1024D duomenų rinkiniui yra *E5* vektorizavimo modelis su reguliarizuotos logistinės regresijos klasifikatoriumi, pritaikytas *Groceries - E-Tech* poroje. Toliau taip pat (žr. 9 lentelė) pateikiamos kitos poros su geriausiu 1024D eksperimentinės grupės metodu, siekiant apžvelgti metodo klasifikavimo gebėjimus, testuojant kitose srityse.

9 lentelė. Likusios poros efektyviausiam metodui (*E5* ir reguliarizuota logistinė regresija) ir geriausiai apmokymo sričiai, *Groceries*, 1024D eksperimentinėje grupėje

Apmokymo-testavimo pora	PRC AUC	ROC AUC
Groceries-E-Niche	0,9954	0,9582
Groceries-Beauty	0,9908	0,9763
Groceries-E-Marketplace	0,9821	0,9375
Groceries-Clothing	0,9757	0,9559
Groceries-Other	0,9483	0,9525

Iš 9 lentelėje pateiktų rezultatų, pastebimas aukštas šio geriausio eksperimentinės grupės metodo, *E5* su reguliarizuota logistine regresija efektyvumas, apmokant su geriausia *Groceries* sritimi, bei testuojant kitose srityse. PRC AUC reikšmės visose porose viršija 0,94 ribą, kas rodo itin aukštą klasifikavimo tikslumą tarp skirtingų mažmeninės prekybos sričių. Verta pažymėti, kad toks rezultatų lygis gali būti siejamas su *E5* vektorizavimo modelio gebėjimu efektyviai perteikti sudėtingus semantinius santykius duomenyse, nepriklausomai nuo apmokymo ar testavimo srities, įskaitant ir išskirtinę, *Other* sritį.

Toliau pateikiama geriausio identifikuoto metodo, *E5* vektorizavimo modelio ir reguliarizuotos logistinės regresijos, *Groceries - E-Tech* poros sumaišymo matrica. Siekiama identifikuoti, kaip geriausia šios eksperimentinės grupės kombinacija prognozuoja klasių žymas, lyginant geriausio metodo rezultatus prognozes bei tikrąsias duomenų rinkinio žymas (žr. 29 pav.).

		Truth data			
		Class 1	Class 2	Classification overall	User's accuracy (Precision)
Classifier results	Class 1	146	9	155	94.194%
	Class 2	191	2347	2538	92.474%
	Truth overall	337	2356	2693	
	Producer's accuracy (Recall)	43.323%	99.618%		
Overall accuracy (OA):		92.573%			
Kappa[1]:		0.559			

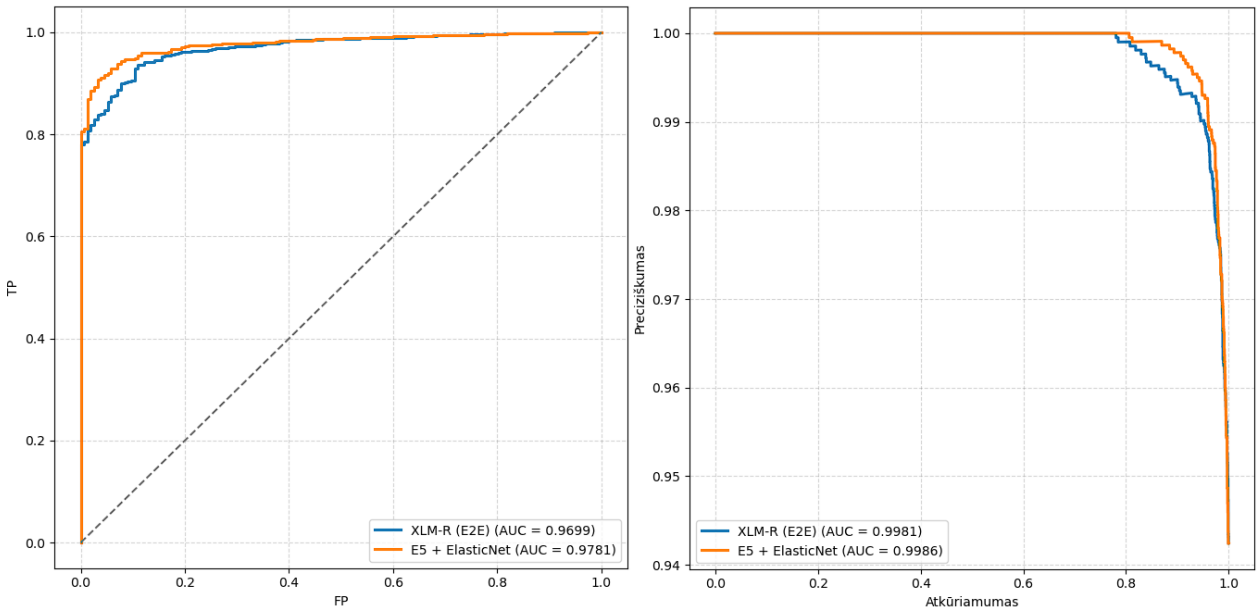
29 pav. 1024D eksperimentinės grupės geriausios kombinacijos sumaišymo matrica

Pateikiamoje sumaišymo matricoje matoma, kad 1024D eksperimentinės grupės geriausiu metodu ir pora pasiekiami aukštesni įverčiai – 92,573 % bendras tikslumas bei vidutinis Kappa koeficiento sutapimo lygis (0,559). Naudojant *E5* ir reguliarizuotą logistinę regresiją šioje tarpsritinėje poroje, taip pat užfiksuojami aukšti rezultatai identifikuojant teigiamą (1) klasę, su 99,618 % preciziškumu ir 92,474 % atkūriamumu. Pastebima, kad metodo efektyvumas yra mažesnis, identifikuojant neigiamas (0) klases, kur atkūriamumo rodiklis siekia 43,323 %. Neigiamos klasės preciziškumas (94,194 %) išlieka aukštas – identifikuotos neigiamos klasės geriausiu metodu dažniausiai būna parenkamos korektiškai.

3.4. Geriausia detekcijos kombinacija

Atlikus eksperimentus 3.3 poskyryje, surastos geriausios tarpsritinės sentimentų analizės metodai ir poros. Vertinant 768D eksperimentinę grupę, tai buvo *XLM-RoBERTa* (*E2E*) klasifikatorius, 1024D atveju – *E5* vektorizavimo modelis su reguliarizuotos logistinės regresijos klasifikatoriumi. Abu šie

metodai identifikuoti kaip geriausi, apmokant metodus su *Groceries*, bei testuojant ant *E-Tech* sričių. Geriausi metodai dar kartą palyginami ROC ir PR kreivėmis (žr. 30 pav.).



30 pav. Geriausių kombinacijų ROC (kairėje) ir PR (dešinėje) kreivės, *Groceries* apmokymo sritis, *E-Tech* testavimo sritis

Pagal PRC AUC ir ROC AUC reikšmes pastebima, jog geriausias 1024D duomenų rinkinio metodas (*E5* + LogReg (*Elastic Net*)) efektyviau prognozuoja sentimentų klases, lyginant su 768D rinkinio geriausiu metodu (*XLM-RoBERTa* (*E2E*)). Tai patvirtinama ir vizualiai, 1024D metodui esant arčiau viršutinio kairio krašto ROC kreivėje bei PR kreivių grafike taip pastebima identiška tendencija – 1024D metodui, lyginant su 768D, pastebimas mažiausias preciziškumo sumažėjimas, didėjant atkūriamumo reikšmėms.

Taip pat svarbu paminėti, kad identifikuotas geriausias *E5* ir reguliarizuotos logistinės regresijos metodas išsiskiria ne tik geriausių rezultatų poroje bet ir kitose porose, kuriose kaip apmokymo sritis naudojama *Groceries*. Visais šiais atvejais ši vektorizacijos ir klasifikavimo kombinacija pateikia aukštus PRC AUC ir ROC AUC įverčius, aplenkiant kitus metodus. Tokie rezultatai leidžia daryti prielaidą, kad *E5* vektorizavimo modeliu išgaunamos plačiai pritaikomos žodžių reprezentacijos, veiksmingos įvairiuose mažmeninės prekybos kontekstuose Lietuvoje. Taip akcentuojamas šios metodikos universalumas ir potencialus tolimesnis pritaikomumas, siekiant identifikuoti patikimus ir efektyvius sentimentų analizės sprendimus lietuvių kalbai.

3.5. Geriausios detekcijos kombinacijos pritaikomumas

Tyrimo metu identifikuota geriausia tarpsritinės sentimentų analizės klasifikavimo metodika – naudojant *E5* vektorizacijos modelį bei reguliarizuotos logistinės regresijos klasifikatorių. Ši kombinacija pasižymi itin aukštu tikslumu identifikuojant skirtingų mažmeninės prekybos sričių atsiliepimų sentimentų klases.

Metodas gali būti pritaikomas mažmeninės prekybos srityje, pasitelkiant šią metodikų kombinaciją dinaminės kainodaros bei reklaminių kampanijų optimizavimo tikslams. Staigūs vartotojų sentimentų pokyčiai, kaip, pavyzdžiui, padidėjęs teigiamų atsiliepimų skaičius apie tam tikrus produktus (pvz.,

dronus, skirtus filmavimui), užfiksuojami, pritaikant šiame darbe identifikuotą geriausią sentimentų analizės metodiką kaip papildomą funkciją į rinkos paklausos prognozavimo modelius. Tokiu būdu pateikiamos įžvalgos leidžia priimti optimalius, bei duomenimis grįstus sprendimus, organizuojant tam tikrų produktų (ar produktų kategorijų) išpardavimus, tuo pačiu metu atitinkamai adaptuojant inventoriaus kiekius. Taip galima užtikrinti patikimesnį bei optimalų vartotojų perteikiamų sentimentų panaudojimą, net per skirtingas mažmeninės prekybos sritis.

Ši metodika taip pat gali būti pritaikoma klientų aptarnavimo srityje. Didelės mažmeninės prekybos įmonės įprastai turi skambučių centrus, kuriuose naudojama specializuota programinė įranga, pavyzdžiui, „Genesys“. Ši platforma pasitelkiama klientų aptarnavimo agentų pokalbiams su vartotojais – skambučio metu, realiu laiku transkribuojami agentų ir klientų pasakomi žodžiai, bei pritaikoma binarinė sentimentų analizė tolimesnėms įžvalgoms, vertinant klientų atsiliepimus apie produktus, įmonę, darbuotojų elgesį, efektyvumą ir t. t. Pritaikant geriausią, modernų tarpsritinės analizės metodą šiame kontekste, potencialiai gali būti gerinama įžvalgų kokybė, leidžianti efektyviau atsižvelgti į vartotojų nuomonę ne vien mažmeninės prekybos, tačiau ir aptarnavimo gerinimo aspektais.

Ilgalaikėje perspektyvoje, ši sentimentų klasifikavimo kombinacija gali būti pritaikoma skirtingų sektorių analizei – identifikuojant vartotojų sentimentų skirtumus tarp skirtingų prekių ar paslaugų kategorijų. Įmonės pritaikydamos šią metodiką, gali įgyti pozicionavimo pranašumą, efektyviai lyginant konkurentų bei savo esamų klientų atsiliepimus bei identifikuojant aiškius privalumus ir trūkumus tam tikroms produktų ar paslaugų sritims. Ši geriausia išrinkta kombinacija taip pat gali būti panaudojama kaip pokalbių robotų (angl. *chatbot*) komponentas, užtikrinant optimalią klientų patirtį, sumažinant klientų praradimo rodiklį (angl. *churn rate*), bei keliant vartotojų suasmeninimo lygį apsipirkimo patirtyse.

Išvados

1. Atlikus literatūros apžvalgą nustatyta, kad sentimentų analizė yra itin vertinga priemonė verslams, siekiant efektyviai analizuoti vartotojų emocinį nusiteikimą bei elgseną šiuolaikiniame skaitmeniniame kontekste. Įmonėms aktualus tampa automatizuotas vartotojų kuriamo turinio apdorojimas, suteikiantis galimybę vertinti klientų atsiliepimus bei efektyviai reaguoti į rinkos pokyčius, taip užtikrinant vartotojų pasitenkinimą bei įgyjant konkurencinį pranašumą. Mažai resursų turinčiose kalbose, kaip lietuvių, reikalingi tinkami analizės metodai, leidžiantys efektyviai apdoroti semantiškai sudėtingas vartotojų žodines išraiškas. Pastaruoju metu sentimentų analizė yra vis plačiau taikoma mažmeninės prekybos srityse, o jos tikslumas tiesiogiai priklauso nuo konkrečios srities ir pasirinktų metodų efektyvumo.
2. Apžvelgti atlikti sentimentų analizės tyrimai Lietuvoje ir nustatyta, kad nors esant dideliam skaičiui darbų, pritaikytų būtent lietuviškam tekstui, naujesniuose tyrimuose remiamasi efektyviais, transformeriais grįstais modeliais bei kitais moderniais metodais. Šios metodikos rodo didesnį potencialą analizuojant sudėtingas pasaulio kalbas, kaip lietuvių. Pastebima, kad apžvelgtuose tyrimuose Lietuvoje, sentimentų analizė dar nebuvo išmėginta tarp skirtingų mažmeninės prekybos sričių.
3. Atliktas turimo duomenų rinkinio papildymas mažmeninių prekybos įmonių atsiliepimais iš „Google Maps“. Rinkinys suskirstytas į 7 atskiras atsiliepimų sritis tolimesniam darbo tyrimui. Išbandyti 5 vektorizavimo bei 4 klasifikavimo metodai, vertinant jų efektyvumą tarpsritiniame sentimentų analizės uždavinyje.
4. Atliktas empirinis tyrimas bei gauti rezultatai parodė modernių vektorizavimo metodų pranašumą šiai sentimentų analizės užduočiai. Eksperimentai, atlikti dviejose skirtingose (768D ir 1024D) dimensijų grupėse parodė, kad efektyviausiai veikė įvairiakalbiams duomenims skirtas, modifikuotas *E5* vektorizavimo modelis, kartu su reguliarizuotos logistinės regresijos klasifikatoriumi. Atrinkus efektyviausią apmokymo sritį – *Groceries*, nustatytos geriausios tarpsritinės poros. Geriausio metodo atveju aukščiausi įverčiai pastebėti, apmokant kombinaciją Groceries srityje, bei testuojant *E-Tech* srityje, fiksuojamas 92,6 % tikslumas (ROC AUC = 0,978, PRC AUC = 0,998). Pastebima, jog dimensionalumo didinimas nežymiai pagerina šios tarpsritinės sentimentų analizės užduoties rezultatus.
5. Tyrimo rezultatai parodė, kad sentimentų analizės metodus galima efektyviai taikyti tarp skirtingų Lietuvos mažmeninės prekybos sričių, pasitelkiant atitinkamai priderintas vektorizavimo ir klasifikavimo metodų kombinacijas. Remiantis atliktu darbu bei geriausiu identifikuotu metodu, pateiktos rekomendacijos tokio tipo sentimentų analizės sistemas pritaikyti verslo praktikoje. Geriausias rastas tarpsritinės sentimentų analizės metodas Lietuvos mažmeninės prekybos kontekste gali būti pritaikomas gerinant vartotojų grįžtamojo ryšio vertinimą, siekiant efektyviai identifikuoti problemines įmonės ar produkto vietas bei norint geriau suprasti vartotojų lūkesčius, optimizuojant klientų aptarnavimo kokybę bei užtikrinant efektyvius, duomenimis grįstus sprendimus.

Literatūros sąrašas

1. KAPOČIŪTĖ-DZIKIENĖ, J., R. DAMAŠEVIČIUS ir M. WOŹNIAK. Sentiment Analysis of Lithuanian Texts Using Traditional and Deep Learning Approaches. *Computers* [interaktyvus]. Basel: MDPI, 2019, vol. 8(1), p. 1-16 [žiūrėta 2025-01-30]. Prieiga per: doi:10.3390/computers8010004.
2. PETKEVIČIUS, M., D. VITKUTĖ-ADŽGAUSKIENĖ ir D. AMILEVIČIUS. Targeted Aspect-Based Sentiment Analysis for Lithuanian Social Media Reviews. Iš: UTKA, A., et al. *Human Language Technologies – The Baltic Perspective* [interaktyvus]. Amsterdam: IOS Press, 2020, p. 32-38 [žiūrėta 2025-01-30]. Prieiga per: doi:10.3233/FAIA200599.
3. LIU, Bing. *Sentiment Analysis and Opinion Mining*. California: Morgan and Claypool Publishers, 2012. ISBN 978-3-031-01017-0.
4. PRÖLLOCHS, N., S. FEUERRIEGEL ir D. NEUMANN. Statistical inferences for polarity identification in natural language. *PLoS ONE* [interaktyvus]. Netherlands: Leiden University, 2018, vol. 13(12) [žiūrėta 2025-01-30]. Prieiga per: doi:10.1371/journal.pone.0209323.
5. THALER, R. H. *Misbehaving: the making of behavioral economics*. New York: W.W. Norton & Company, 2015. ISBN 978-0-393-24677-3.
6. ASCHEMANN-WITZEL, J. ir S. ZIELKE. Can't Buy Me Green? A Review of Consumer Perceptions of and Behavior Toward the Price of Organic Food. *Journal of Consumer Affairs* [interaktyvus]. New Jersey: Wiley, 2017. vol. 51(1), p. 211-251 [žiūrėta 2025-01-30]. Prieiga per: doi:10.1111/joca.12092.
7. MERBAH, N. ir S. BENITO-HERNÁNDEZ. Consumer Willingness-to-Pay for Sustainable Coffee: Evidence from a Choice Experiment on Fairtrade and UTZ Certification. *Sustainability* [interaktyvus]. Basel, Switzerland: MDPI, 2024. vol. 16, no. 8, p. 1-12 [žiūrėta 2025-01-30]. Prieiga per: doi:10.3390/su16083222.
8. TIMOSHENKO, A. ir J. R. HAUSER. Identifying Customer Needs from User-Generated Content. *Marketing Science* [interaktyvus]. Maryland, USA: Marketing Science, INFORMS, 2019. vol. 38, no. 1, p. 1-20 [žiūrėta 2025-01-30]. Prieiga per: doi:10.1287/mksc.2018.1123.
9. MÄNTYLÄ, M. V., D. GRAZIOTIN ir M. KUUTILA. The Evolution of Sentiment Analysis - A Review of Research Topics, Venues, and Top Cited Papers. *Computer Science Review* [interaktyvus]. Amsterdam: Elsevier, 2018, vol. 27, p. 16-32 [žiūrėta 2025-01-30]. Prieiga per: doi:10.1016/j.cosrev.2017.10.002.
10. TURNEY, P. D. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* [interaktyvus]. Philadelphia: Association for Computational Linguistics, 2002, p. 417-424 [žiūrėta 2025-01-30]. Prieiga per: doi:10.48550/arXiv.cs/0212032.
11. CAMBRIA, E., B. SCHULLER, Y. XIA ir C. HAVASI. New Avenues in Opinion Mining and Sentiment Analysis. *IEEE Intelligent Systems* [interaktyvus]. New Jersey: IEEE, 2013. vol. 28, no. 2, p. 15-21 [žiūrėta 2025-01-30]. Prieiga per: doi:10.1109/MIS.2013.30.
12. ZHANG, L., S. WANG ir B. LIU. Deep Learning for Sentiment Analysis: A Survey. *WIREs Data Mining and Knowledge Discovery* [interaktyvus]. New Jersey: Wiley, 2018 [žiūrėta 2025-01-30]. Prieiga per: doi:10.1002/widm.1253.
13. PANG, B. ir L. LEE. *Opinion mining and sentiment analysis*. Boston; Delft: Now Publishers, 2008. ISBN 978-1-60198-150-9.

14. MINGHUA, N. ir G. CHAOFAN. Hybrid of Spans and Table-Filling for Aspect-Level Sentiment Triplet Extraction. Iš: CALZOLARI, A., et al. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* [interaktyvus]. Torino: ELRA & ICCL, 2024. p. 8464-8473 [žiūrėta 2025-01-30]. ISBN 978-2-493814-10-4.
15. ZHENG, J. ir Y. LIU. Probing Language Identity Encoded in Pre-rained Multilingual Models: A Typological View. *PeerJ Computer Science* [interaktyvus]. London: PeerJ, 2022. vol. 8, e899 [žiūrėta 2025-01-30]. Prieiga per: doi:10.7717/peerj-cs.899.
16. UYSAL, A. K. ir S. GÜNAL. The Impact of Preprocessing on Text Classification. *Information Processing & Management* [interaktyvus]. Amsterdam: Elsevier, 2014. vol. 50, no. 1, p. 104-112 [žiūrėta 2025-01-30]. Prieiga per: doi:10.1016/j.ipm.2013.08.006.
17. ZHU, L. ir D. LUO. A Novel Efficient and Effective Preprocessing Algorithm for Text Classification. *Journal of Computer and Communications* [interaktyvus]. California: Scientific Research Publishing, 2023, vol. 11, no. 3, p. 1-14 [žiūrėta 2025-01-30]. Prieiga per: doi:10.4236/jcc.2023.113001.
18. VERMA, B. ir R. S. THAKUR. Sentiment Analysis Using Lexicon and Machine Learning-Based Approaches: A Survey. Iš: TIWARI, B., et. al. *Proceedings of International Conference on Recent Advancement on Computer and Communication* [interaktyvus]. Singapore: Springer Singapore, 2018. p. 441-447 [žiūrėta 2025-01-30]. Prieiga per: doi:10.1007/978-981-10-8198-9_46.
19. SENNRICH, R., B. HADDOW ir A. BIRCH. Neural Machine Translation of Rare Words with Subword Units. Iš: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* [interaktyvus]. Philadelphia: Association for Computational Linguistics, 2016, p. 1715-1725 [žiūrėta 2025-01-30]. Prieiga per: doi:10.18653/v1/P16-1162.
20. KAPOČIŪTĖ-DZIKIENĖ, J., A. KRUPAVIČIUS ir T. KRILAVIČIUS. A Comparison of Aproaches for Sentiment Classification on Lithuanian Internet Comments. Iš: *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing* [interaktyvus]. Philadelphia: Association for Computational Linguistics, 2013. p. 2-11 [žiūrėta 2025-01-30]. ISBN 9781937284596.
21. MANNING, C. D., P. RAGHAVAN ir H. SCHÜTZE. *Introduction to information retrieval*. New York: Cambridge University Press, 2008. ISBN 978-0-521-86571-5.
22. ABUBAKAR, H. D., M. UMAR ir M. A. BAKALE. Sentiment Classification: Review of Text Vectorization Methods: Bag of Words, Tf-Idf, Word2vec and Doc2vec. *SLU Journal of Science and Technology* [interaktyvus]. Nigeria: Sule Lamido University, 2022, vol. 4(1-2), p. 27–33 [žiūrėta 2025-01-30]. Prieiga per: doi:10.56471/slujst.v4i.266.
23. KRZESZEWSKA, U., A. PONISZEWSKA-MARAŃDA ir J. OCHELSKA-MIERZEJEWSKA. Systematic Comparison of Vectorization Methods in Classification Context. *Applied Sciences* [interaktyvus]. Basel: MDPI, 2022, vol. 12(10), p. 5119 [žiūrėta 2025-01-30]. Prieiga per: doi:10.3390/app12105119.
24. ANILSAGAR, T. ir S. S. A. SYED. The Evolution of Sentiment Analysis and Conversational AI: Techniques, Applications, and Future Research Directions. *Journal of Systems Engineering and Electronics* [interaktyvus]. 2025, vol. 35(1), p. 71-82 [žiūrėta 2025-01-30]. ISSN 1671-1793.
25. AGUSTINA, C., P. PURWANTO ir F. FARIKHIN. Enhancing Sentiment Analysis Accuracy in Borobudur Temple Visitor Reviews through Semi-Supervised Learning and SMOTE Upsampling. *Journal of Advances in Information Technology* [interaktyvus]. California:

Engineering and Technology Publishing, 2024, vol. 15(4), p. 492–499 [žiūrėta 2025-01-30]. Prieiga per: doi:10.12720/jait.15.4.492-499.

26. RAHMAN, Md. M., A. I. SHIPLU, Y. WATANOBE ir Md. A. ALAM. RoBERTa-BiLSTM: A Context-Aware Hybrid Model for Sentiment Analysis [interaktyvus]. 2024, 18 p. [žiūrėta 2025-01-30]. Prieiga per: doi:10.48550/arXiv.2406.00367.
27. KULKARNI, Jay Milind. Sentiment Analysis of Hindi Song Lyrics using a BiLSTM Model with BERT Embeddings: magistro baigiamasis darbas [interaktyvus]. National College of Ireland, 2023 [žiūrėta 2025-01-30]. Prieiga per: https://norma.ncirl.ie/7208/
28. MISHEV, K., et al. Evaluation of Sentiment Analysis in Finance: From Lexicons to Transformers. *IEEE Access* [interaktyvus]. New Jersey: IEEE, 2020, vol. 8, 131662–131682 [žiūrėta 2025-01-30]. Prieiga per: doi:10.1109/ACCESS.2020.3009626.
29. VASWANI, A., et al. Attention Is All You Need. *31st Conference on Neural Information Processing Systems (NIPS 2017)* [interaktyvus]. 2017, 15 p. [žiūrėta 2025-01-30]. Prieiga per internetą: http://arxiv.org/abs/1706.03762.
30. ZHANG, H., ir M. O. SHAFIQ. Survey of transformers and towards ensemble learning using transformers for natural language processing. *Journal of Big Data* [interaktyvus]. London: SpringerOpen, 2024, vol. 11(1), 25 [žiūrėta 2025-01-30]. Prieiga per: doi:10.1186/s40537-023-00842-0.
31. HASHMI, E., S. Y. YAYILGAN ir S. SHAIKH. Augmenting sentiment prediction capabilities for code-mixed tweets with multilingual transformers. *Social Network Analysis and Mining* [interaktyvus]. Berlin, Springer Nature 2024, vol. 14(1), 86 [žiūrėta 2025-01-30]. Prieiga per: doi:10.1007/s13278-024-01245-6.
32. SANH, V., L. DEBUT, J. CHAUMOND ir T. WOLF. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter [interaktyvus]. 2019 [žiūrėta 2025-01-30]. Prieiga per: doi:10.48550/arXiv.1910.01108.
33. AWLLA, K. M., H. VEISI ir A. A. ABDULLAH. Sentiment analysis in low-resource contexts: BERT‘s impact on Central Kurdish [interaktyvus]. Netherlands: Springer, 2025, vol. 1, 34 [žiūrėta 2025-01-30]. Prieiga per: doi:10.1007/s10579-024-09805-0.
34. DEVLIN, J., M. W. CHANG, K. LEE ir K. TOUTANOVA. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [interaktyvus]. 2019 [žiūrėta 2025-01-30]. Prieiga per: doi:10.48550/arXiv.1810.04805.
35. SHAH, S. M. A. H., et al. Arabic Sentiment Analysis and Sarcasm Detection Using Probabilistic Projections-Based Variational Switch Transformer. *IEEE Access* [interaktyvus]. New Jersey: IEEE, 2023, vol. 11, 67865–67881 [žiūrėta 2025-01-30]. Prieiga per: doi:10.1109/ACCESS.2023.3289715.
36. ZHEO, S., et al. Multi-source multi-modal domain adaptation. *Information Fusion* [interaktyvus]. Amsterdam: Elsevier, 2025, vol. 117 [žiūrėta 2025-02-17]. Prieiga per: doi:10.1016/j.inffus.2024.102862.
37. BEN-DAVID, S., J. BLITZER, K. CRAMMER ir F. PEREIRA. Analysis of Representations for Domain Adaptation. Iš: SCHÖLKOPF, B., et al. *Advances in Neural Information Processing Systems 19*. Cambridge: The MIT Press, 2007. p. 137–144 [žiūrėta 2025-02-17]. ISBN 978-0-262-25691-9.

38. PAN, S. J., I. W. TSANG, J. T. KWOK ir Q. YANG. Domain Adaptation via Transfer Component Analysis. *IEEE Transactions on Neural Networks* [interaktyvus]. New Jersey: IEEE, 2011, vol. 22(2), p. 199–210 [žiūrėta 2025-02-17]. Prieiga per: doi:10.1109/TNN.2010.2091281.
39. FARAHANI, A., S. VOGHOEI, K. RASHEED ir H. R. ARABNIA. A Brief Review of Domain Adaptation. Iš: STAHLBOCK, R., et. al. *Advances in Data Science and Information Engineering* [interaktyvus]. Cham: Springer International Publishing, 2021, p. 877–894 [žiūrėta 2025-02-17]. Prieiga per: doi:10.1007/978-3-030-71704-9_65.
40. GANIN, Y., et al. Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning 17* [interaktyvus]. New York: JMLR, 2016, vol. 17(59), p. 1-35 [žiūrėta 2025-02-17]. Prieiga per: doi:10.48550/arXiv.1505.07818.
41. YANG, Y., J. KO ir S. Y. YUN. Towards Difficulty-Agnostic Efficient Transfer Learning for Vision-Language Models. *Proceddings of the 2024 Conference on Empirical Methods in Natural Language Processing* [interaktyvus]. Florida: Association for Computational Linguistics, 2024, p. 2066-2085. [žiūrėta 2025-02-17]. Prieiga per: doi:10.18653/v1/2024.emnlp-main.124.
42. RICHTER-PECHANSKI, P., et al. Clinical information extraction for lower-resource languages and domains with few-shot learning using pretrained language models and prompting. Iš: MITKOV, R. *Natural Language Processing* [interaktyvus]. New York: Cambridge University Press, 2024, p. 1-24 [žiūrėta 2025-02-17].Prieiga per: doi:10.1017/nlp.2024.52.
43. AMBRAZAS, V. *Lithuanian grammar*. 2-as pataisytas leidimas. Vilnius: Baltos Lankos, 2006. ISBN 9955-23-035-5.
44. VILEIKYTĖ, B., M. LUKOŠEVIČIUS ir L. STANKEVIČIUS. Sentiment analysis of Lithuanian online reviews using large language models. *Information society and university studies* [interaktyvus]. Kaunas: Kaunas University of Technology, 2024 [žiūrėta 2025-02-23]. Prieiga per: doi:10.48550/arXiv.2407.19914.
45. ŠTRIMAITIS, R., P. STEFANOVIČ, S. RAMANAUSKAITĖ, A. SLOTKIENĖ. Financial Context News Sentiment Analysis for the Lithuanian Language. *Applied Sciences* [interaktyvus]. Basel: MDPI, 2021, vol. 11(10), p. 4443 [žiūrėta 2025-02-23]. Prieiga per doi:10.3390/app11104443.
46. MORKŪNAITĖ, Laura. Sentimento poliariškumo tyrimas Lietuvos įmonių klientų atsiliepimuose veidaknygėje ir evertink.lt: magistro baigiamasis darbas [interaktyvus]. Kauno technologijos universitetas, 2019 [žiūrėta 2025-02-23]. Prieiga per: https://epubl.ktu.edu/object/elaba:37753631
47. DAUGĖLA, Kęstutis. E. verslo paslaugų vartotojų lietuviškų atsiliepimų klasifikavimas: magistro baigiamasis darbas [interaktyvus]. Kauno technologijos universitetas, 2018 [žiūrėta 2025-02-23]. Prieiga per: https://epubl.ktu.edu/object/elaba:28994167/
48. ASHBAUGH, L. ir Y. ZHANG. A Comparative Study of Sentiment Analysis on Customer Reviews Using Machine Learning and Deep Learning. *Computers* [interaktyvus]. 2024, vol. 13(12), 340 [žiūrėta 2025-02-23]. Prieiga per: doi:10.3390/computers13120340.
49. WILIANTO T., S. SUPRYADI ir A. WIBOWO. Sentiment Analysis on E-commerce Product using Machine Learning and Combination of TF-IDF and Backward Elimination. *International Journal of Recent Technology and Engineering* [interaktyvus]. 2020, vol. 8(6), p. 2862–2867 [žiūrėta 2025-02-23]. Prieiga per: doi:10.35940/ijrte.F7889.038620.

50. WANG, Zekai. Sentiment Analysis of Mobile Phone Reviews Using XGBoost and Word Vectors. *ITM Web of Conferences* [interaktyvus]. 2025. vol. 70 [žiūrėta 2025-02-23]. Prieiga per: doi:10.1051/itmconf/20257003018.
51. PRYTULA, Marianna. Fine-tuning BERT, DistilBERT, XLM-RoBERTa and Ukr-RoBERTa models for sentiment analysis of ukrainian language reviews. *Artificial Intelligence* [interaktyvus]. 2024. vol. 29(2), p. 85–97 [žiūrėta 2025-02-23]. Prieiga per: doi: 10.15407/jai2024.02.085.
52. KOTLER, P. ir K. L. KELLER. *Marketing management*. 15-as leidimas. India: Pearson India Education, 2015. ISBN 978-93-325-5718-5.
53. STEINBERGER, J. ir K. JEZEK. Using Latent Semantic Analysis in Text Summarization and Summary Evaluation. Iš: BENEŠ M. *Proceddings of the 7th International Conference on Information Systems Implementation and Modelling* [interaktyvus]. Czechia: MARQ, 2004. p. 93-100 [žiūrėta 2025-02-23]. Prieiga per: http://textmining.zcu.cz/publications/isim.pdf
54. YAHAV, I., O. SHEHORY ir D. SCHWARTZ. Comments Mining With TF-IDF: The Inherent Bias and Its Removal. *IEEE Transactions on Knowledge and Data Engineering* [interaktyvus]. 2019. vol. 31(3), p. 437–450 [žiūrėta 2025-03-02]. Prieiga per: doi: doi:10.1109/TKDE.2018.2840127.
55. STURUA, S., et al. jina-embeddings-v3: Multilingual Embeddings With Task LoRA [interaktyvus]. 2024, 20 p. [žiūrėta 2025-03-02]. Prieiga per: doi:10.48550/arXiv.2409.10173.
56. WANG, L., et al. Text Embeddings by Weakly-Supervised Contrastive Pre-training [interaktyvus]. 2022, 17 p. [žiūrėta 2025-03-02]. Prieiga per: doi:10.48550/arXiv.2212.03533.
57. PHUOC, Chu Duong Huy. FuocChuVIP123 at CoMeDi Shared Task: Disagreement Ranking with XLM-Roberta Sentence Embeddings and Deep Neural Regression [interaktyvus]. 2025, 6 p. [žiūrėta 2025-03-02]. Prieiga per: doi:10.48550/arXiv.2501.12336.
58. CONNEAU, A., et al. Unsupervised Cross-lingual Representation Learning at Scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* [interaktyvus]. Florida: Association for Computational Linguistics, 2020, p. 8440-8451 [žiūrėta 2025-03-02]. Prieiga per: doi:10.18653/v1/2020.acl-main.747.
59. PAN, S. J. ir Q. YANG. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering* [interaktyvus]. 2009. vol. 22(10), p. 1345–1359 [žiūrėta 2025-03-10]. Prieiga per: doi:10.1109/TKDE.2009.191.
60. ZOU, H. ir T. HASTIE. Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* [interaktyvus]. New Jersey: Wiley, 2005, vol. 67(5), p. 301-320 [žiūrėta 2025-03-10]. Prieiga per: doi:10.1111/j.1467-9868.2005.00503.x
61. CORTES, C. Ir V. VAPNIK. Support-Vector Networks. *Machine Learning* [interaktyvus]. Boston: Kluwer Academic Publishers, 1995, 20, p. 273–297 [žiūrėta 2025-03-10]. Prieiga per: doi:10.1007/BF00994018.
62. HEARST, M. A., et al. Support vector machines. *IEEE Intelligent Systems and their Applications* [interaktyvus]. New Jersey: IEEE, 1998, Vol. 13(4), p. 18–28 [žiūrėta 2025-03-10]. Prieiga per: doi:10.1109/5254.708428.
63. JOACHIMS, Thorsten. Text Categorization with Support Vector Machines [interaktyvus]. Germany: University of Dortmund, 1999 [žiūrėta 2025-03-10]. Prieiga per: doi:10.17877/DE290R-5097.

64. GEURTS, P., D. ERNST ir L. WEHENKEL. Extremely randomized trees. *Machine Learning* [interaktyvus]. Berlin: Springer Science + Business Media, LLC, 2006. vol. 63(1), p. 3–42 [žiūrėta 2025-03-10]. Prieiga per: doi:10.1007/s10994-006-6226-1.
65. BEGER, Andreas. Precision-Recall Curves. *SSRN Electronic Journal* [interaktyvus]. Amsterdam: Elsevier, 2016, 7 p. [žiūrėta 2025-03-20]. Prieiga per: doi:10.2139/ssrn.2765419.
66. BREIMAN, Leo. Random Forests. *Machine Learning* [interaktyvus]. Berlin: Springer Science + Business Media, LLC, 2001, Vol. 45, p. 5–32 [žiūrėta 2025-03-20]. Prieiga per: doi:10.1023/A:1010933404324.
67. LIU, X. Q., Q. L. WU ir W. T. PAN. Sentiment classification of micro-blog comments based on Randomforest algorithm. *Concurrency and Computation: Practice and Experience* [interaktyvus]. New Jersey: Wiley, 2019, vol. 31(10) [žiūrėta 2025-03-20]. Prieiga per: doi:10.1002/cpe.4746.
68. HAIZHOU, L., et al. Baselines and Bigrams: Simple, Good Sentiment and Topic Classification. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* [interaktyvus]. South Korea: Association for Computational Linguistics, 2012, p. 90–94 [žiūrėta 2025-03-23]. Prieiga per: https://aclanthology.org/P12-2018/.
69. WHITEHEAD, M. ir L. YAEGER. Building a General Purpose Cross-Domain Sentiment Mining Model. Iš: *2009 WRI World Congress on Computer Science and Information Engineering* [interaktyvus]. California: IEEE, 2009. p. 472–476 [žiūrėta 2025-03-23]. Prieiga per: doi: 10.1109/CSIE.2009.754.
70. DEERWESTER, S., et al. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* [interaktyvus]. 1990, vol. 41(6), p. 391–407 [žiūrėta 2025-03-28]. Prieiga per: doi:10.1002/(SICI)1097-4571(199009)41:63.0.CO;2-9.
71. AHMED, A. Z. Ir M. RODRÍGUEZ-DÍAZ. Significant Labels in Sentiment Analysis of Online Customer Reviews of Airlines. *Sustainability* [interaktyvus]. Basel, Switzerland: MDPI, 2020. vol. 12(20), 8683 [žiūrėta 2025-03-28]. Prieiga per: doi:10.3390/su12208683.
72. LI, Z., et al. Towards General Text Embeddings with Multi-stage Contrastive Learning. [interaktyvus]. 2023, 18 p. [žiūrėta 2025-04-05]. Prieiga per: doi:10.48550/arXiv.2308.03281.
73. ZHANG, X., et al. mGTE: Generalized Long-Context Text Representation and Reranking Models for Multilingual Text Retrieval. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track* [interaktyvus]. Florida: Association for Computational Linguistics, 2024. p. 1393–1412. [žiūrėta 2025-04-09]. Prieiga per: https://aclanthology.org/2024.emnlp-industry.103/
74. WANG, L., et al. Multilingual E5 Text Embeddings: A Technical Report [interaktyvus]. 2024, 6 p. [žiūrėta 2025-05-11]. Prieiga per: https://arxiv.org/abs/2402.05672.