Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Klientų praradimo mažinimas telekomunikacijų paslaugų sektoriuje, naudojant skaitmeninių platformų duomenis

Baigiamasis magistro studijų projektas

Paulius Žilinskas

Projekto autorius

Prof. dr. Jūratė Banytė

Vadovė

Doc. dr. Loreta Saunorienė

Vadovė

Kaunas, 2024

Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Klientų praradimo mažinimas telekomunikacijų paslaugų sektoriuje, naudojant skaitmeninių platformų duomenis

Baigiamasis magistro studijų projektas

Didžiųjų verslo duomenų analitika (6213AX001)

Paulius Žilinskas

Projekto autorius

Prof. dr. Jūratė Banytė
Vadovė

Doc. dr. Loreta Saunorienė
Vadovė

Doc. dr. Asta Tarutė
Recenzentė

Vyr. lekt. dr. Vilma Petrauskienė
Recenzentė

Kaunas, 2024

Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Paulius Žilinskas

Klientų praradimo mažinimas telekomunikacijų paslaugų sektoriuje, naudojant skaitmeninių platformų duomenis

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.

Paulius Žilinskas

Patvirtinta elektroniniu būdu

Paulius Žilinskas. Klientų praradimo mažinimas telekomunikacijų paslaugų sektoriuje, naudojant skaitmeninių platformų duomenis. Magistro studijų baigiamasis projektas / vadovė prof. dr. Jūratė Banytė, vadovė doc. dr. Loreta Saunorienė; Kauno technologijos universitetas, Matematikos ir gamtos mokslų fakultetas.

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

Reikšminiai žodžiai: klientų praradimas, telekomunikacijų paslaugos, sentimentų analizė, mašininis mokymasis, natūralios kalbos apdorojimas, temos modeliavimas, atsiliepimai.

Kaunas, 2024. 63 p.

Santrauka

Šiandieniniame verslo pasaulyje klientų išlaikymo strategija yra būtina, siekiant sėkmingai konkuruoti rinkoje. Didžioji dalis įmonių skiria daug pastangų bei laiko analizuodamos įvairius duomenis ir tokiu būdu stengiasi užkirsti kelią klientų praradimui. Sudaromi įvairūs modeliai, kurie pagal kliento elgseną gali nuspėti jo praradimą, tačiau pieš sudarant tokius modelius būtina žinoti, kokios priežastys dažniausiai paskatina klientą atsisakyti naudojamų paslaugų ir pereiti pas konkurentus. Šis klausimas ypač aktualus greita dinamika pasižyminčiame telekomunikacijų paslaugų sektoriuje. Šio projekto tikslas – naudojant skirtingus skaitmeninėse platformose paskelbtų klientų atsiliepimų tyrybos metodus, identifikuoti klientų praradimą telekomunikacijų paslaugų sektoriuje lemiančius veiksnius ir pateikti gautais rezultatais pagrįstus jo mažinimo sprendimus.

Literatūros analizės metu identifikuoti klientų praradimą telekomunikacijų paslaugų sektoriuje lemiantys veiksniai, kuriuos autoriai įvardija kaip pagrindinius. Remiantis kitų autorių tyrimų rezultatais, sudarytas konceptualusis modelis, kurį sudaro šie veiksniai: kaina, ryšio kokybė, aptarnavimas, reklama ir savitarnos sistema. Tuomet atlikta įvairių matematinių metodų apžvalga, kurie naudojami natūralios kalbos apdorojime.

Šiame projekte pasirinkta analizuoti viešai prieinamus atsiliepimus apie 3 didžiausias telekomunikacijų paslaugų įmones Lietuvoje. Atsiliepimų vektorizavimui panaudoti 4 skirtingi metodai: žodžių krepšelis, termino dažnis – atvirkštinis dokumento dažnis, Word2Vec (skip-gram) ir Word2Vec (CBOW). Rankiniu būdu priskirtiems sentimentams klasifikuoti išbandyti 5 skirtingi metodai: gradiento didinimas, XGBoost, atsitiktinis miškas, logistinė regresija ir k-artimiausių kaimynų. Iš sudarytų 20 klasifikavimo modelių geriausią rezultatą pavyko pasiekti naudojant logistinės regresijos modelį ir termino dažnio – atvirkštinio dokumento dažnio vektorizavimo metodą. Vėliau, neigiamiems atsiliepimams buvo atliktas temos modeliavimas, pasinaudojant LDA metodu. Paskutiniame rezultatų poskyryje naudotas sudarytas Word2Vec (skip-gram) modelis bei analizuotas neigiamų atsiliepimų kontekstas.

Tyrimo metu nustatyta, kad klientų praradimą telekomunikacijų paslaugų sektoriuje labiausiai veikia trys veiksniai: paslaugų kaina, ryšio kokybė bei įrenginių remontas. Pirmieji du veiksniai buvo identifikuoti teorinėje dalyje, kaip turintys didelę įtaką klientų praradimui. O įrenginių remonto veiksnys išryškėjo tik tyrimo dalyje.

Žilinskas, Paulius. Reducing Customer Churn in the Telecommunications Services Sector Using Data from Digital Platforms. Master's Final Degree Project / supervisors prof. Jūratė Banytė and assoc. prof. Loreta Saunorienė; Faculty of Mathematics and Natural Sciences, Kaunas University of Technology.

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

Keywords: customer churn, telecommunications services, sentiment analysis, machine learning, natural language processing, topic modeling, reviews.

Kaunas, 2024. 63 pages.

Summary

In today's fast-paced business world, a customer retention strategy is essential for successfully competing in the market. The majority of companies invest significant effort and time analyzing various data in order to prevent customer loss. Various models are developed to predict customer churn based on their behavior, but in creating such models, it is crucial to understand the common reasons that prompt customers to abandon the services they use and switch to competitors. This question is particularly relevant in the rapidly evolving telecommunications sector. The objective of this project is to identify the factors leading to customer churn in the telecommunications sector by utilizing various research methods based on customer feedback published on different digital platforms, and to present evidence-based solutions for reducing customer churn based on the obtained results.

During the analysis of literature, factors leading to customer churn in the telecommunications sector were identified by the authors as key. Based on the results of studies by other authors, a conceptual model was constructed, in which the following factors were identified: price, quality of service, customer support, advertising, and self-service system. Subsequently, a review of various mathematical methods used in natural language processing was conducted.

In this study, I chose to analyze publicly available reviews about the top 3 telecommunications companies in Lithuania. Four different methods were used for vectorizing the reviews: bag of words (BOW), term frequency-inverse document frequency (TF-IDF), Word2Vec (skip-gram), and Word2Vec (CBOW). Five different methods were tested for sentiment classification: gradient boosting, XGBoost, random forest, logistic regression, and k-nearest neighbors. Among the 20 classification models created, the best result was achieved using the logistic regression model and the TF-IDF vectorization method. Subsequently, topic modeling was performed on negative reviews using the LDA method. In the final section of the results, the Word2Vec (skip-gram) model was utilized, and the context of negative reviews was analyzed.

During the research, it was found that the primary factors influencing customer churn in the telecommunications sector are service price, quality of communication, and device repairs. The first two factors were identified in the theoretical part as having a significant impact on customer churn. Meanwhile, the factor of device repairs was only discovered in the practical part of the study.

Turinys

Lentelių sąrašas

Paveikslų sąrašas

Įvadas

Tyrimo aktualumas. Kiekviena paslaugų įmonė, norinti išlikti rinkoje ir įgyti konkurencinį pranašumą, privalo žinoti galimas klientų praradimo priežastis ir nuolat jas analizuoti. Tai leidžia priimti savalaikius, klientų praradimo mažinimą įgalinančius sprendimus. Paskutiniaisiais metais pastebimas ženklus mokslinių straipsnių, kuriuose klientų praradimas analizuojamas panaudojant mašininio mokymosi metodus, augimas [19]. Įmonės tokiu būdu stengiasi neatsilikti nuo tendencijų ir vis dažniau siekia kurti prognozavimo modelius, kurie gebėtų identifikuoti besitraukiančius klientus. Šis būdas veiksmingas lėtos dinamikos įmonėse, tačiau telekomunikacijų paslaugų sektorius tuo nepasižymi. Tai reiškia, kad klientai paprastai per trumpą laiko tarpą pereina iš vieno paslaugų teikėjo pas kitą [19]. Telekomunikacijų paslaugų sektoriuje siekiant kuo tikslesnių klientų praradimo prognozių, būtina tiksliai žinoti veiksnius, kurie lemia jų sprendimą atsisakyti naudojamų paslaugų. Kitu atveju labai tikėtina, kad apie galimą kliento praradimą modelis perspės gerokai per vėlai.

Dalis klientų, norėdami išreikšti nepasitenkinimą arba padėką, palieka atsiliepimus skaitmeninėje erdvėje. Interneto svetainėse arba kitose skaitmeninėse platformose parašyti atsiliepimai įmonėms gali suteikti naudingos informacijos apie kritines ar mažiau reikšmingas klientų praradimo priežastis, tačiau būtina šiuos duomenis tinkamai apdoroti, kad rezultatai būtų tikslūs ir lengvai interpretuojami.

Tyrimo objektas. Klientų praradimą telekomunikacijų paslaugų sektoriuje lemiantys veiksniai.

Tyrimo tikslas. Naudojant skirtingus skaitmeninėse platformose paskelbtų klientų atsiliepimų tyrybos metodus, identifikuoti klientų praradimą telekomunikacijų paslaugų sektoriuje lemiančius veiksnius ir pateikti gautais rezultatais pagrįstus jo mažinimo sprendimus.

Tyrimo uždaviniai:

1. Atskleisti klientų praradimo sampratos esmę, tipologijas ir klientų praradimo valdymo ypatumus telekomunikacijų paslaugų sektoriuje.
2. Sudaryti konceptualųjį klientų praradimą telekomunikacijų paslaugų sektoriuje lemiančių veiksnių modelį, pagrįstą mokslinių tyrimų rezultatais.
3. Apžvelgti skaitmeninėse platformose paskelbtų klientų atsiliepimų tyrybai tinkamus metodus, susijusius su natūralios kalbos apdorojimu.
4. Naudojant surinktus klientų atsiliepimus, atlikti sentimentų priskyrimą ir tuomet išbandyti skirtingus vektorizavimo ir klasifikavimo metodus.
5. Neigiamiems klientų atsiliepiamas atlikti temos modeliavimą bei konteksto analizę.
6. Apibendrinti skirtingų klientų atsiliepimų tyrybos metodų taikymu pagrįstus tyrimų rezultatus ir pateikti įžvalgas klientų praradimui telekomunikacijų paslaugų sektoriuje mažinti.

Tyrimo metodai: mokslinės literatūros analizė, klasifikavimas, teksto vektorizavimas, temos modeliavimas, konteksto analizė, vizualizavimas.

1. Literatūros analizė

1.1. Klientų išlaikymo svarba ir tyrimų apžvalga

Klientų išlaikymas yra vienas iš esminių tvarius įmonių rezultatus lemiančių veiksnių. Išlaikyti esamus klientus yra daug pigiau ir lengviau nei surasti naujus, ypač brandžiose rinkose [1]. Amoako, Arthur‘o, Bandoh ir Katah [2] atliktas tyrimas parodė, kad tarp esamų klientų išlaikymo ir naujų pritraukimo yra reikšmingas skirtumas. Taip yra todėl, kad naujų klientų pritraukimo išlaidos paprastai yra nuo penkių iki net dvidešimt penkių kartų didesnės nei esamų klientų išlaikymo. Verta atkreipti dėmesį ne tik į didelę naujų klientų pritraukimo kainą, bet ir į įmonių nesugebėjimą išlaikyti esamus klientus. Naujausiuose tyrimuose pastebima, kad šiandienėje neramioje ir perpildytoje verslo aplinkoje klientų išlaikymas vis dar yra didelis iššūkis daugeliui įmonių ir teigiama, kad ateityje reikės taikyti patobulintus klientų išlaikymo metodus. Kitu atveju įmonės bus priversto ieškoti naujų klientų ir patirs kur kas didesnes finansines išlaidas [3].

Klimavičienė ir Lingaitienė [4] teigia, kad norint turėti lojalų klientą, reikia išlaikyti glaudžius santykius su juo. Palaikant pastovų ryšį su klientu vyksta lengvesnis tarpusavio bendravimas bei supratimas, tai padeda nuspėti poreikius ar planus ir užtikrinti ilgalaikį bendradarbiavimą. Šiam tikslui pasiekti įmonės įgyvendina į klientus orientuotas strategijas, diegdamos santykių su klientais valdymo sistemas (angl. *Customer relationship management, CRM*). CRM sistemos naudą klientų lojalumui atskleidžia Keropyan‘o ir Gil-Lafuente [5] tyrimo rezultatai. Jais remiantis konstatuojama, kad CRM sistema leidžia įmonėms valdyti rinkodaros strategijas ir teikti konkrečias paslaugas skirtingų vertybių klientams. Tai padeda išlaikyti pelningiausius klientus ir didinti jų lojalumą.

Bankų sektoriuje atlikti tyrimai parodė, kad siekiant išlaikyti esamus klientus, pagrindinis vaidmuo atitenka esamiems darbuotojams. Tai reiškia, kad bankai privalo turėti profesionaliai išmokytus darbuotojus bei nuolat tobulinti jų gebėjimus, kad klientai jautųsi saugūs ir negalvotų apie banko keitimą [6]. Huarng‘as ir Hui-Kuang‘as [7] analizavo kainų didėjimo poveikį klientų išlaikymui. Tyrimo metu pastebėta, kad kainų pokytis turi įtakos klientų pasitenkinimui, o pasitenkinimas veikia klientų išlaikymą. Lyginant lojalius ir nelojalius klientus pastebėta, kad lojalūs klientai yra tolerantiškesni padidėjusioms kainoms. O nelojalūs klientai yra kur kas jautresni didėjančioms kainoms. Ieškant informacijos apie ryšį tarp klientų pasitenkinimo ir klientų išlaikymo, galima rasti skirtingų nuomonių. Dažniausiai teigiama, kad pasitenkinimas visada teigiamai veikia išlaikymą. Tačiau naujesniuose tyrimuose pabrėžiama, kad klientų pasitenkinimas gali nukristi iki neigiamo, bet išlaikymas tuo metu gali padidėti iki neutralaus ar net teigiamo. Tai reiškia, kad klientų pasitenkinimo sumažėjimas nebūtinai turi lemti ir klientų išlaikymo rodiklio kritimą.

Mahmoud‘o, Hinson‘o, Adika‘o [8] tyrimas, kuris atliktas Ganos telekomunikacijų srityje, nagrinėjo pasitikėjimo, įsipareigojimo ir konfliktų sprendimo įtaką klientų išlaikymui, įtraukiant ir kliento pasitenkinimo veiksnį. Šio tyrimo išvados parodė, kad pasitikėjimas ir konfliktų sprendimas yra reikšmingai susiję su klientų pasitenkinimu. Taip pat nustatyta, kad pasitikėjimas ir konfliktų sprendimas turi netiesioginį ryšį su klientų išlaikymu per klientų pasitenkinimą. Tyrimo išvadose pripažįstama, kad klientų išlaikymui įtakos turi klientų pasitenkinimo lygis, todėl įmonės privalo orientuotis į klientų pasitenkinimo gerinimą, nes tai visada padidina klientų išlaikymą. Konfliktų sprendimas buvo nustatytas kaip vienas iš svarbiausių rodiklių, analizuojant klientų pasitenkinimą ir išlaikymą. Tuo remiantis daroma išvada, kad mobiliojo ryšio paslaugų teikėjams ypač svarbu sukurti ir įdiegti visapusišką ir veiksmingą konfliktų ir skundų valdymo sistemą. Be to, paslaugų teikėjai

turėtų dėti reikiamas pastangas, kad klientų nusiskundimai būtų išspręsti operatyviai, o jei nepavyksta į juos sureaguoti iš karto, informuoti klientus, kada planuojama juos išspręsti.

Jau minėtame Keropyan‘o ir Gil-Lafuente [5] tyrime taip pat buvo siekiama analizuoti ir skirtingoms telekomunikacijų paslaugų klientų grupėms pritaikyti jų lojalumui stiprinti skirtas šešias programas:

1. Apdovanojimų programa. Mobilieji telefonai ir planšetiniai kompiuteriai pradžiugintų daugelį klientų, tačiau ši programa priskiriama tik vienam klientų segmentui, kurį sudaro aukštesnės klasės klientai. Klientai iš aukštesnio socialinio sluoksnio, kurie turi didesnes pajamas, dažniausiai negaili pinigų įvairioms paslaugoms. Todėl tikimasi, kad vertingos dovanos pagerins jų lojalumą ir tuo pačiu atneš įmonei finansinės naudos.
2. Premijų programa. Ji apima taškus, nemokamus kuponus ir mažesnius skambučių įkainius. Antroji lojalumo programa pritaikyta žemesnes ir vidutines pajamas gaunantiems klientams. Ši programa gali būti interpretuojama taip, kad antrasis segmentas labiausiai vertina dovanų kuponus arba siūlomus žaidimus, kurie leistų kaupti taškus ir tuomet įsigyti pigesnius arba papildomų naudų teikiančius planus.
3. Sumažintos savaitgalio kainos programa. Ši programa priskiriama segmentui, apimančiam klientus, gyvenančius miestuose.
4. Nemokamų SMS programa. Priskiriama segmentams, apimantiems paauglius, namų šeimininkes ir vyresnio amžiaus žmones.
5. Nemokamų mobiliųjų programų programa. Ši programa taip pat turėtų būti priskirta jauniems klientams, kurie daugiau laiko praleidžia naudodamiesi įvairiomis mobiliosiomis programomis.
6. Dviejų valandų nemokamo mobiliojo interneto programa. Priemiesčiuose gyvenantys klientai paprastai turi mažesnes pajamas, todėl jiems gali patikti nemokama mobiliojo interneto paslauga.

Daroma išvada, kad aptartos programos leidžia pagrįstai suskirstyti klientus į grupes ir pasiūlyti jiems priimtiniausią variantą.

Atlikta mokslinės literatūros analizė atskleidžia, kad klientų išlaikymo strategija yra būtina kiekvienai įmonei, siekiančiai sėkmingai konkuruoti rinkoje. Santykių su klientais sistemų diegimas, darbuotojų ugdymas, teisingų kainų nustatymas, efektyvus konfliktų sprendimas bei lojalumo programos yra svarbiausi veiksniai, lemiantys gerus klientų išlaikymo rezultatus. Be to, nustatyta, kad klientų pasitenkinimas yra labai glaudžiai susijęs su klientų išlaikymu, o tai reiškia, kad patenkinti klientai kur kas rečiau pagalvoja apie įmonės palikimą. Kartu pripažįstant, kad visų esamų įmonės klientų išlaikymas yra neįmanomas, svarbu pasigilinti ir į klientų praradimo reiškinį ir galimus jo būdus.

1.2. Klientų praradimo samprata, tipologijos ir ypatumai paslaugų sektoriuje

Kaip teigia Seo, Ranganathan‘as ir Yair‘as [9], klientų praradimas yra rimta nesėkmė įmonei, kalbant apie dabartinį ir būsimą pelną. Tai situacija, kai kliento indėlis į įmonės pajamas mažėja [10]. Skačkauskienės ir Toropovaitės [11] atliktame tyrime lojalių klientų praradimas apibūdinamas ne tik kaip pardavimų sumažėjimą sukeliantis veiksnys. Tai reiškia viso pardavimo srauto, kurį vartotojas būtų sukūręs per bendradarbiavimo su įmone laiką, praradimą. Trumpiausiai ir aiškiausiai klientų praradimą apibūdina Lazarov‘as ir Capota [12]. Jie šį reiškinį apibūdino kaip kliento apsisprendimą atsisakyti paslaugų, produktų ar net pačios įmonės ir pereiti pas konkurentą.

Shobana, Gangadhar‘as, Arora, Renjith‘as, Bamini ir Chincholkar‘as [10] teigia, kad klientų praradimas gali būti skirstomas į dvi grupes pagal tai, ar įmonė ir klientas yra sudarę sutartį. Sutarties

sudarymo situacijoje parodoma, kad įmonė ir klientas bendradarbiauja, siekdami sumažinti abiejų šalių galimas patirti rizikas. Pasirašius susitarimo dokumentą klientas privalo laikytis įsipareigojimų, kurie aprašyti sutartyje. Jei sąlygų nėra laikomasi, tuomet numatomos tam tikros sankcijos, dažniausiai klientas privalo padengti įmonei patirtus nuostolius. Šiuo atveju yra labai lengva identifikuoti paslaugų sutartį nutraukusį ir įmonę palikusį klientą. Tuo tarpu įmonės ir kliento santykiuose, kuriuose nėra pasirašoma sutartis, klientas gali bet kada pasirinkti konkurento teikiamas paslaugas ar prekes. Tai reiškia, kad klientas gali pirkti įmonės produktą ilgą laiką arba pirkti pirmą ir paskutinį kartą. Šiuo atveju identifikuoti prarastą ar dar neprarastą klientą yra gana sudėtinga.

Klientai, apsiperkantys internetu, dažniausiai yra priskiriami nesutartiniams santykiams. Nesutartinių klientų praradimas gali būti skirstomas į dvi kategorijas [10]:

- **Laikinas praradimas**. Tai reiškia klientus, kurie per tam tikrą laiką neįsigijo įmonės prekių ar paslaugų. Bet tai nereiškia, kad klientas yra prarastas, jis bet kada gali vėl įsigyti įmonės prekę ar paslaugą;
- **Visiškas praradimas**. Šis terminas reiškia kliento sprendimą ateityje nepirkti paslaugų ar įmonės produktų. Šį kliento apsisprendimą gali lemti įvairios priežastys, įskaitant kliento pirkimo įpročių pasikeitimą, taip pat augimo stadijos pasikeitimą, kai produktas nebereikalingas.

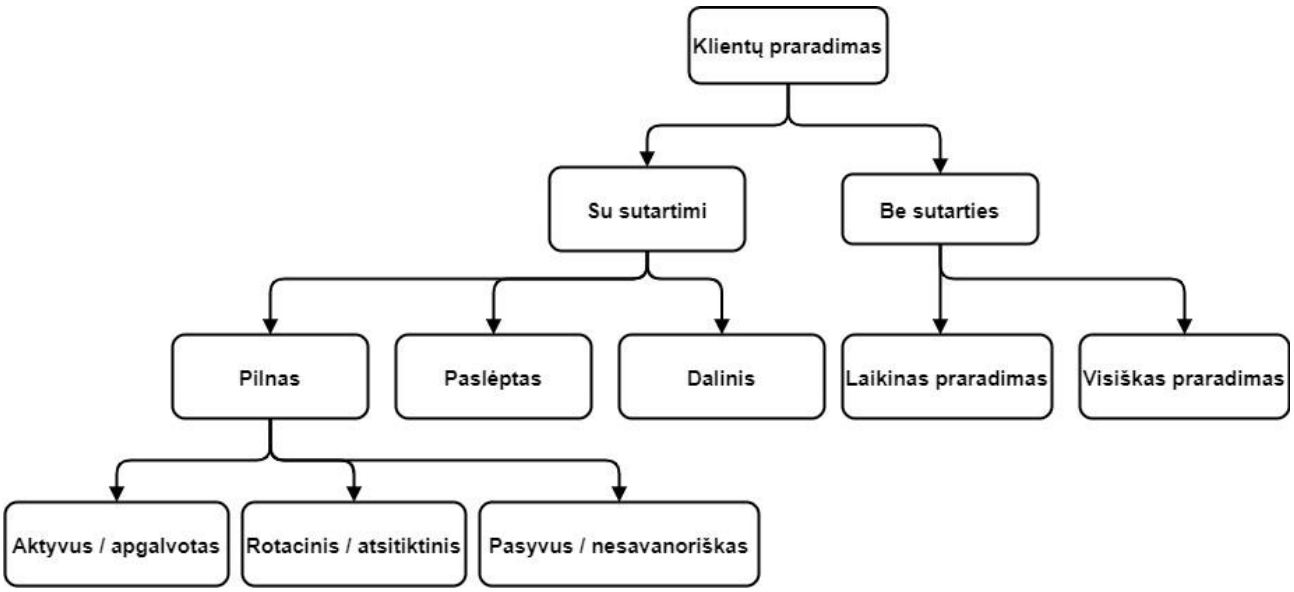
Dažnai yra galvojama, kad sutartinis klientas yra prarandamas tik tada, kai yra oficialiai nutraukiama sutartis. Tačiau tai nėra tiesa, kliento pasitraukimą galima atpažinti ir iš jo elgsenos dar prieš sutarties pabaigą ar nutraukimą. Kliento praradimas gali būti 3 tipų [12]:

- **Pilnas** – sutartis oficialiai nutraukta;
- **Paslėptas** – sutartis nenutraukiama, tačiau klientas jau seniai aktyviai nesinaudoja paslaugomis;
- **Dalinis** – sutartis nėra nutraukta ir klientas iš dalies naudojasi teikiamomis paslaugomis, tačiau tuo pat metu naudojamos ir konkurentų paslaugos.

Be to, literatūroje galima rasti 3 sutartinių klientų pasitraukimo tipus nutraukiant sutartį [12]:

- **Aktyvus / apgalvotas** – klientas nusprendžia nutraukti sudarytą sutartį ir pereiti pas konkurentą. Priežastys gali būti įvairios: nepasitenkinimas paslaugos kokybe, nekonkurencinga kaina, įmonės nesirūpinimas lojaliais klientais ir pan.;
- **Rotacinis / atsitiktinis** – klientas nutraukia sutartį be tikslo pereiti pas konkurentą. Dažniausiai tai lemia aplinkybių pasikeitimai, kurie neleidžia klientui toliau naudotis paslauga, pvz. finansinės problemos, dėl kurių neįmanoma apmokėti sąskaitos, arba kliento geografinės padėties pasikeitimas į vietą, kurioje įmonės nėra arba paslauga nepasiekiama;
- **Pasyvus / nesavanoriškas** – įmonė pati nutraukia sutartis.

Visi anksčiau išvardinti klientų praradimo tipai pavaizduoti 1.1 paveiksle.



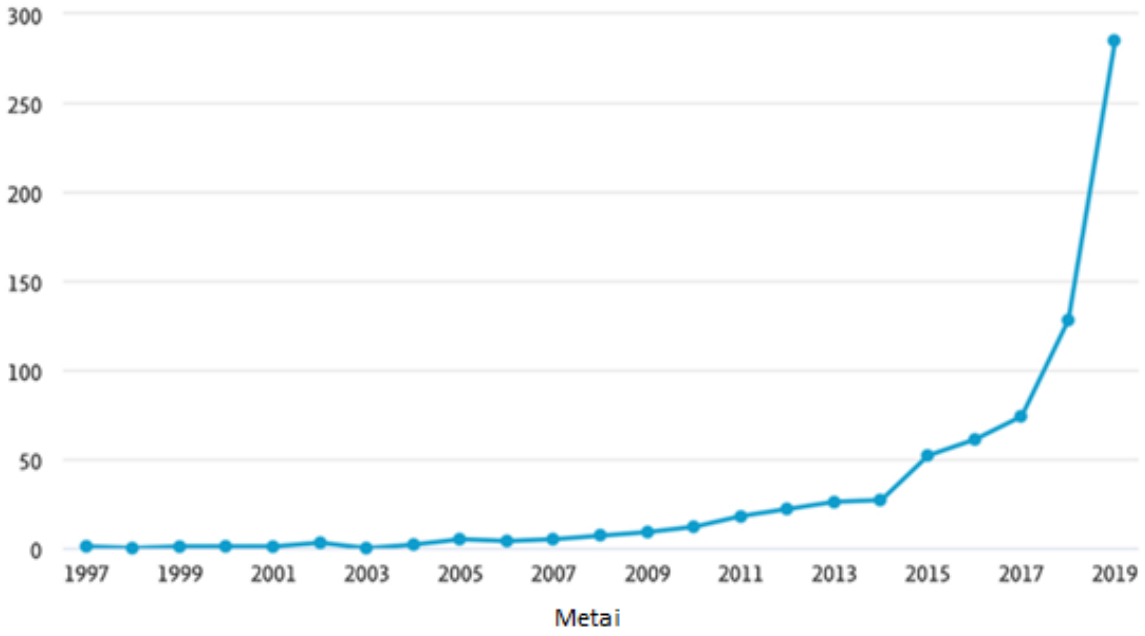
1.1 pav. Klientų praradimo tipai (sudaryta pagal Shobana, Gangadhar'ą, Arora, Renjith'ą, Bamini ir Chincholkar'ą [10] bei Lazarov'ą ir Capota [12])

Aptariant išskirtus klientų praradimo tipus, pažymėtina, kad įmonė, norėdama išvengti didelių finansinių nuostolių, turėtų identifikuoti ir išsamiau vertinti paslėptą ir dalinį klientų praradimo tipus (pvz., klientas moka mėnesinį abonentinį mokestį, bet visiškai nesinaudoja gaunamomis paslaugomis). Identifikavus tokį klientą galima daryti prielaidą, kad jis planuoja atsisakyti įmonės teikiamų paslaugų ir pereiti pas konkurentus, kurie siūlo patrauklesnę lojalumo programą ar kitus privalumus [12]. Kita vertus, dažnai teigiama, kad paslaugų sektoriui paprastai būdingi ilgalaikiai klientų ir įmonių santykiai. Nors lojalumo lygiai skiriasi priklausomai nuo paslaugos tipo, klientai tradiciškai naudojasi ir pasitiki vienu paslaugos teikėju ir jį keičia retai.

Kartu reikia pripažinti, kad skaitmenizacijos procesai užtikrina ir palengvina klientų prieigą prie svarbios informacijos apie jų pasirinktas įmones, naudojamas paslaugas bei konkurentų pasiūlymus. Tai didina klientų praradimo grėsmę ir skatina klientų praradimo priežastims [14] identifikuoti skirtų mokslinių tyrimų plėtotę. Bhattacharyya ir Dash'as [16] atliko 211 mokslinių darbų analizę, kuri parodė, kad didžioji dalis mokslininkų, tiriančių klientų praradimą, yra iš JAV ir Kinijos. Klientų praradimo reiškinio pažinimą įgalinantys tyrimai analizuoti ir kitame straipsnyje [19], kuriame pastebėtas sparčiai augantis susidomėjimas klientų praradimo tema pasitelkiant mašininį mokymąsi (1.2 pav.). Tam greičiausiai įtakos turi padidėjęs duomenų prieinamumas, nes įmonės daugiau investuoja į didžiųjų duomenų sprendimus, bei klientų praradimo klausimo svarbos supratimas ir susirūpinimas juo.

Lima'os Lemos, Silva'os ir Tabak'o [19] atliktame tyrime teigiama, kad sparti technologijų pažanga, globalizacija ir „fintech" atsiradimas ženkliai padidino konkurenciją bankų sektoriuje iki dar nematyto lygio. Tam įtakos turėjo mobiliųjų technologijų ir socialinių tinklų plitimas, išaugusi vartotojų prieiga prie informacijos, pagerėjęs klientų finansinis raštingumas, galimybė bendrauti su žmonėmis iš bet kurios pasaulio vietos. Nemaža dalis klientų pradėjo naudoti užsienio įmonių teikiamas paslaugas, kurios yra prieinamos jų regione. Užsienio įmonių teikiamos paslaugos padidino klientų lūkesčius ir tai sumažino jų lojalumą įmonėms, kuriomis jie naudojasi jau ilgą laiką. Todėl nemaža dalis klientų pradėjo galvoti apie teikiamų paslaugų atsisakymą. Didėjanti klientų praradimo grėsmė kelia didelį susirūpinimą įvairioms paslaugų įmonėms ir tai tampa vis rimtesne problema.

Atliktas tyrimas mobiliųjų telekomunikacijų paslaugų sektoriuje parodė, kad klientų praradimo rodiklis svyruoja nuo 20 % iki 40 %. Skaičiuojama, kad sumažinus klientų praradimo rodiklį 5 %, pelnas galėtų padidėti nuo 25 % iki 85 % [13].



1.2 pav. Publikacijų skaičius mokslo žurnaluose ir konferencijose, kuriuose analizuojamas mašininis mokymasis ir klientų praradimas [19]

Apibendrinant išanalizuotą mokslinę literatūrą, galima teigti, kad klientų praradimas yra gana plati sąvoka. Neužtenka vien identifikuoti kliento praradimo, būtina rasti ir tikslią priežastį, kad būtų galima imtis reikiamų veiksmų ir užkirsti tam kelią. Klientų praradimo tema ypač aktuali skaitmeninėje aplinkoje, kurioje klientai gali labai lengvai rasti informaciją apie konkurentus ir jų teikiamas paslaugas bei pasirinkti tai, kas jiems yra priimtiniau. Siekiant teorinio galimų klientų praradimo šiandieniniame paslaugų sektoriuje priežasčių pagrindimo, svarbu analizuoti klientų praradimo ir jo valdymo modelius.

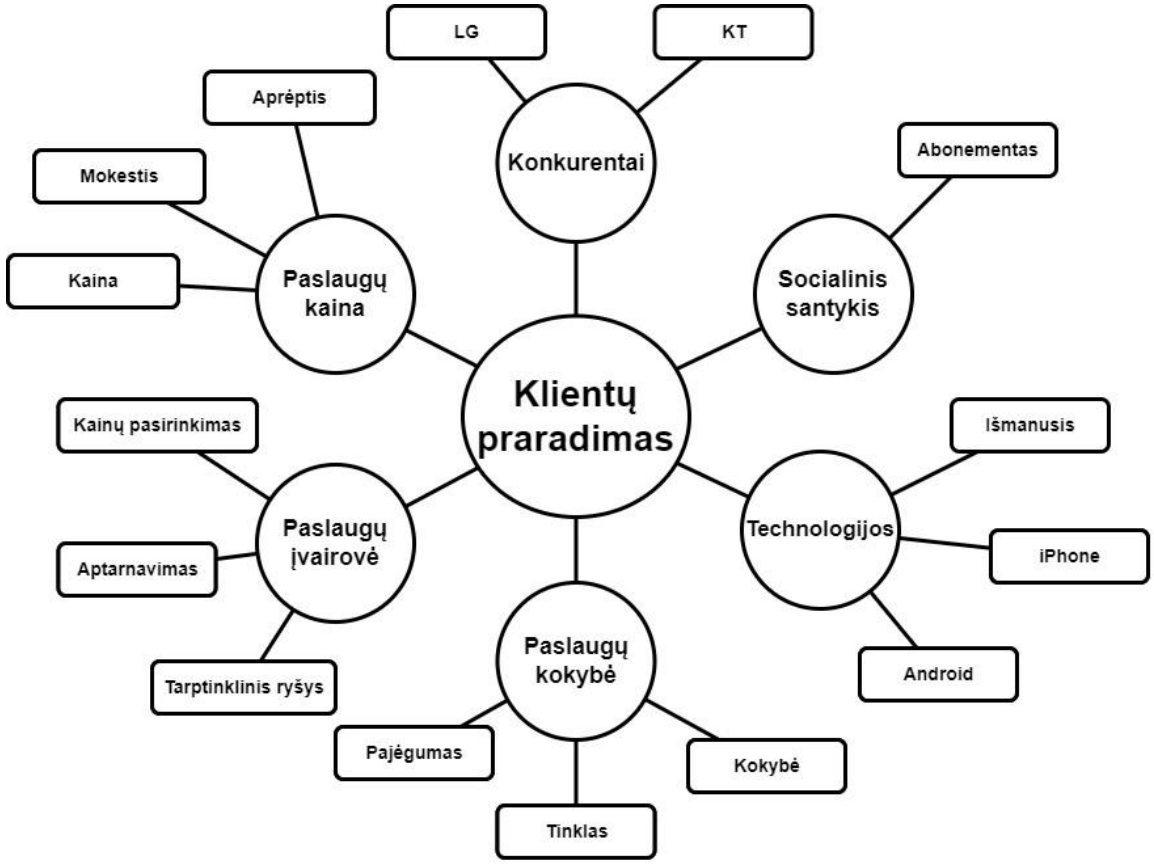
1.3. Klientų praradimo valdymo modeliai ir jų taikymas telekomunikacijų paslaugų sektoriuje

Jau minėtame Lima‘os Lemos, Silva‘os ir Tabak‘o [19] tyrime išskiriami penki principai, kuriuos būtina žinoti, norint sėkmingai valdyti klientų praradimą bei išlaikymą:

1. Sėkmingas klientų išlaikymas sumažina poreikį ieškoti naujų klientų, todėl įmonės gali sutelkti didesnį dėmesį į santykių su esamais klientais stiprinimą. Tai reiškia, kad tinkamas klientų praradimo valdymo modelis leidžia sutaupyti ir tuo pat metu stiprinti esamų klientų lojalumą.
2. Lojalūs klientai, ilgiau besinaudojantys įmonės paslaugomis, yra linkę daugiau išleisti. Taip pat jie nemažai bendrauja bei yra linkę daugiau dalintis teigiamomis ir neigiamomis patirtimis su aplinkiniais.
3. Ilgalaikių klientų aptarnavimas ir išlaikymas yra pigesnis dėl jų jau turimų geresnių žinių, įgytų per jų ilgametę patirtį, naudojantis įmonės paslaugomis.
4. Ilgalaikiai klientai paprastai yra mažiau imlūs konkurentų pasiūlymams.
5. Klientų praradimas yra rimta problema, nes tai sumažina pardavimus, todėl būtina kuo skubiau rasti naujų klientų, kad būtų kompensuoti nuostoliai.

Be to, teigiama, kad yra labai svarbu vertinti sprendimo nutraukti santykius laiko dinamiką. Klientas dažniausiai galvoja apie santykių nutraukimą kelis mėnesius prieš tai padarydamas. Bankų sektorius pasižymi lėta dinamika, todėl prognozuoti kliento netekimą yra gana nesunku. Vienas iš greitos dinamikos pavyzdžių yra telekomunikacijų paslaugų sektorius. Jam būdinga tai, kad klientai paprastai per trumpą laiko tarpą pereina iš vieno operatoriaus pas kitą, todėl prognozavimas yra gana sudėtinga užduotis. Kita vertus, aklai pasitikėti vien laiko dinamika negalima, nes nustatyta, kad atitinkami pokyčiai ekonomikoje, verslo sutrikimai ar net politinė krizė gali turėti įtakos klientų polinkiui palikti paslaugų įmones per labai trumpą laikotarpį.

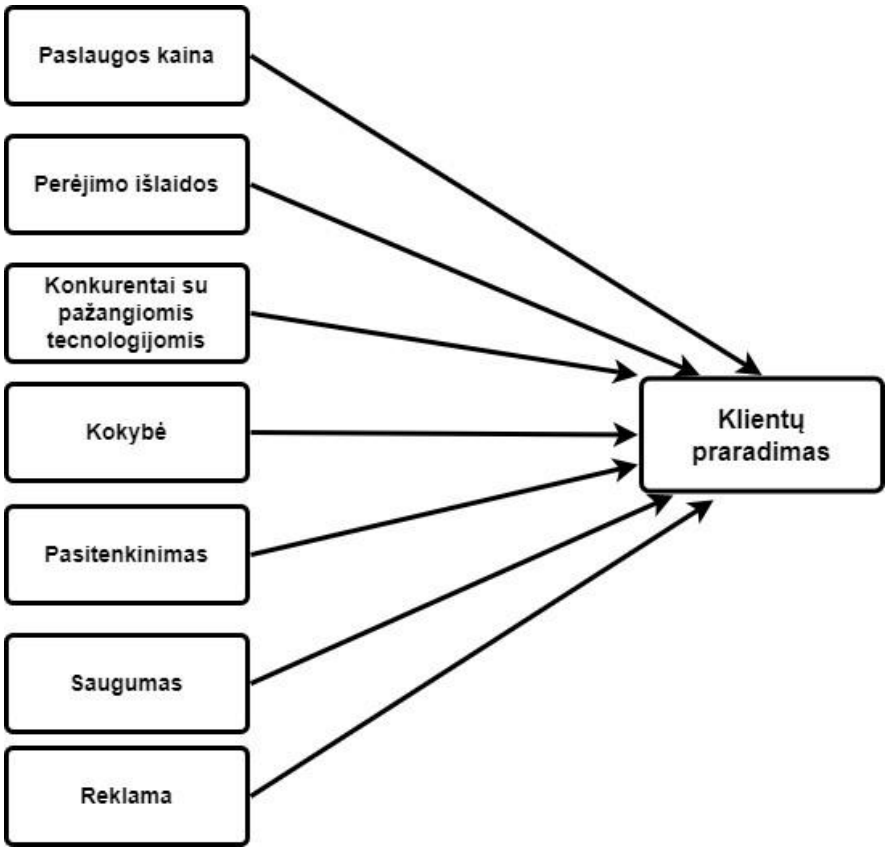
Lee ir kitų autorių tyrime [15] analizuojamas Korėjos telekomunikacijų paslaugų sektorius ir siekiama suprasti, kaip klientų praradimo dažnis priklauso nuo internetinėje žiniasklaidoje skelbiamų žodžių. Atliekant tyrimą buvo surinkti žodžiai, kurie, kaip nustatyta ankstesnėse studijose, turėjo įtakos klientų praradimui. Po to šie žodžiai buvo suskirstyti į kategorijas, siekiant ištirti jų poveikį klientų praradimui (1.3 pav.).



1.3 pav. Klientų praradimo tyrimo modelis (Lee, Kim ir Lee [15])

Iš 1.3 paveiksle pateikto tyrimo modelio matyti, kad tiriant klientų praradimą Korėjos telekomunikacijų paslaugų sektoriuje, išskirtos 5 su klientų praradimo tema susijusios kategorijos bei joms priskirti žodžiai. Po tris žodžius atiteko paslaugų įvairovei, kainai ir kokybei. Technologijų kategorija taip pat gavo 3 žodžius. Tuo tarpu konkurentų kategorijai priskirti 2 žodžiai, o socialinių santykių tik 1 žodis.

Hejazinia ir Kazemi‘io [18] tyrime identifikuoti 7 veiksniai, turintys įtakos klientų praradimui telekomunikacijų paslaugų sektoriuje (1.4 pav.).



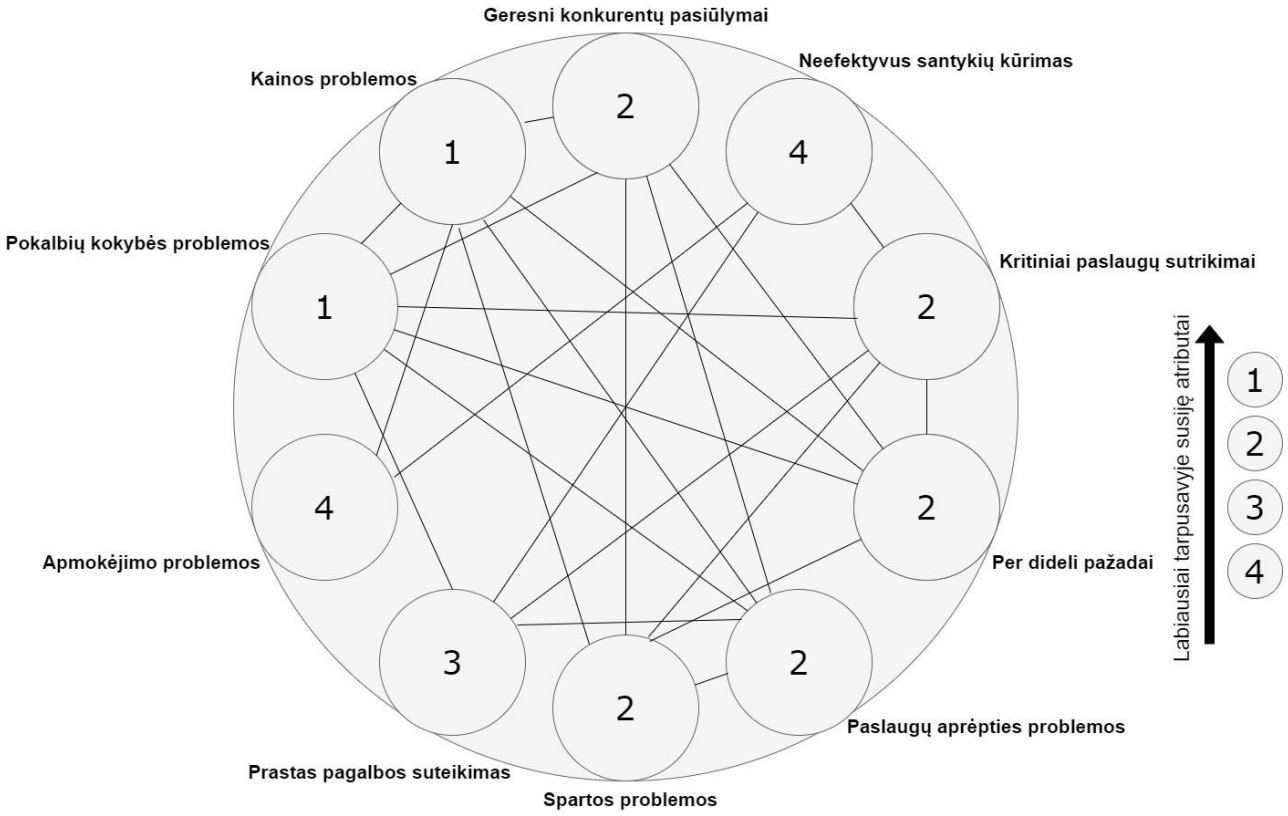
1.4 pav. Klientų praradimo modelis (Hejazinia ir Kazemi [18])

1.4 paveiksle pavaizduoti Hejazinia ir Kazemi‘io [18] išskirti veiksniai interpretuojami taip:

1. **Paslaugos kaina**. Paslaugos kaina reiškia pinigų sumą, kurią klientas turėtų sumokėti už paslaugos suteikimą. Pripažįstama, kad klientai ieško naujų rinkų, kuriose paslaugų kainos būtų mažesnės. Dėl šios priežasties esami paslaugų teikėjai stengiasi neatsilikti nuo konkurentų ir pasiūlyti patrauklius pasiūlymus, kad sumažintų klientų praradimo rodiklį bei pritrauktų daugiau naujų klientų. Daugelyje tyrimų nustatyta, kad didesnės kainos turi neigiamą poveikį klientų pritraukimui ir teigiamą poveikį klientų praradimui.
2. **Perėjimo išlaidos**. Perėjimo išlaidos apibrėžiamos kaip išlaidos, atsirandančios, kai klientai pasirenka konkurentų teikiamas paslaugas. Šios išlaidos gali būti ne tik finansinės, bet ir fizinės ar emocinės. Tokiais atvejais dažnai klientai netenka kai kurių privilegijų ir ypatingų galimybių, kurios jiems buvo suteiktos kaip lojaliems klientams. Taigi, jei klientai susiduria su didelėmis perėjimo išlaidomis, jie nenori keisti paslaugų teikėjo. Galima daryti išvadą, kad perėjimo išlaidos turi neigiamą poveikį klientų apsisprendimui keisti paslaugų teikėją.
3. **Konkurentai su pažangiomis technologijomis**. Konkurentai su naujausiomis technologijomis ir siūloma mažesne kaina nesunkiai sugeba pritraukti naujus klientus. Pavyzdžiui, konkurento teikiamos spartesnės paslaugos, yra didelis pavojus kiekvienai įmonei. Kitaip tariant, jei pasitenkinimas esamomis paslaugomis mažėja, klientai noriai pereina pas kita paslaugų teikėją, siūlantį pažangesnes technologijas.

4. **Kokybė**. Kliento suvokiama paslaugų kokybe dažniausiai įvardijamas skirtumas tarp paslaugų teikėjo pažadėtų ir faktiškai gaunamų paslaugų. Todėl paslaugų įmonėse kokybė priklauso nuo klientų lūkesčių ir įmonės teikiamų paslaugų profesionalumo.
5. **Pasitenkinimas**. Klientų pasitenkinimas reiškia kliento suvoktą vertę atėmus lūkesčius. Kitaip tariant, jei klientas mano, kad gauta vertė atitinka jo lūkesčius, sukuriamas pasitenkinimas. Kitas apibrėžimas įvardija klientų pasitenkinimą kaip bendrą klientų požiūrį į prekę / paslaugą ja pasinaudojus Pasitenkinimas padeda klientui likti su įmone ir tokiu būdu užkerta kelią kliento praradimui.
6. **Saugumas**. Susirūpinimas dėl saugumo reiškia baimę prarasti duomenis ar asmeninę informaciją dėl įmonės kaltės. Pavyzdžiui, telefono pokalbių klausymasis ar asmeninės informacijos teikimas tretiesiems asmenims. Nesugebėjimas užtikrinti saugumo gali turėti įtakos klientų praradimui.
7. **Reklama**. Vienas iš reklamos tikslų yra sukurti stiprų klientų pasitikėjimą įmone. Todėl reklama padeda įmonėms pritraukti naujus klientus, stiprinti esamų lojalumą ir užkirsti kelią klientų mažėjimui.

Bhattacharyya ir Dash‘o tyrime [16], atspindinčiame klientų praradimo tematiką, taip pat pasirenkamas telekomunikacijų paslaugų sektorius. Remiantis internetinių bendruomenių diskusijomis, buvo sudarytas modelis, kuriame pavaizduoti visi klientų praradimą galintys lemti veiksniai ir jų tarpusavio ryšiai. (1.5 pav.).



1.5 pav. Klientų mažėjimo priežasčių ir jų tarpusavio ryšių modelis (Bhattacharyya ir Dash [16])

Šiame modelyje vaizduojami ryšiai tarp įvairių telekomunikacijų paslaugų atributų, kurie gali lemti klientų praradimą. Tokie atributai kaip kaina ir skambučių kokybė yra labiausiai susiję su kitais tyrimo atributais. Pavyzdžiui, klientai yra nepatenkinti gaunamų paslaugų kokybės ir kainos santykiu. Taip gali nutikti dėl to, kad jie yra nusivylę duomenų perdavimo sparta, kurios tikėjosi pasirinkdami

4G ryšį, ir jiems reikia mokėti paslaugų mokestį, kuris yra didesnis nei daugelio konkurentų. Taigi, šios priežastys yra tarpusavyje labai susijusios, todėl vienos iš šių eliminavimas, labai tikėtina, išspręstų ir kitas problemas [16].

Lamrhari'io ir kitų autorių [18] atliktame tyrime teigiama, kad įmonės pirmiausia turėtų atkreipti dėmesį į skaitmeninėse platformose pateikiamus duomenis, norėdamos suprasti savo klientus, ir tik tuomet formuluoti veiksmingas rinkodaros strategijas. Norint turėti stiprią rinkodaros strategiją, reikia suprasti visus klientų tipus: potencialus, naujas, esamas, besiruošiantis palikti arba jau palikęs organizaciją. Pavyzdžiui, pasisveikinimo programa naudojama potencialiems klientams pritraukti, naujai prisijungusiems klientams siunčiamos individualizuotos paslaugos ir pažintiniai pasiūlymai. Pasitenkinimo lygio įvertinimas ir lojalumo programų tobulinimas padeda išlaikyti esamus klientus ir taip pat paversti juos aktyviais ir pelną nešančiais [17].

Didžioji dalis literatūros, susijusios su klientų praradimu, daugiausia dėmesio skiria klientų praradimo prognozavimui. Ypač pastaruoju metu įmonės yra linkusios investuoti į klientų praradimo prognozavimą, nes tai padeda įvertinti, kurie klientai artimiausiu metu planuoja atsisakyti įmonės teikiamų paslaugų. Prognozavimui yra naudojamas statistinis (esamų istorinių duomenų) modeliavimas, kad būtų sukurta balų sistema, pagal kurią būtų galima numatyti esamų klientų praradimą [14].

Tuo tarpu nagrinėjant sentimentus skaitmeninėse platformose, reikėtų pabrėžti, kad tyrimas neatskleis konkrečių klientų, kurie planuoja palikti įmonę, tačiau šio tyrimo rezultatai suteikia galimybę susipažinti su kintamaisiais, kurie gali lemti klientų mažėjimą [14].

Tradicinis statistinis modeliavimas prognozuojant klientų praradimą dažniausiai naudoja tokius modelius kaip logistinė regresija (angl. *Logistic regression*), išgyvenimo modeliai (angl. *Survival models*), neuroniniai tinklai (angl. *Neural networks*) ir savaime besitvarkantys žemėlapiai (angl. *Self-organising maps*). Visiems šiems modeliams reikalingi egzistuojančių klientų duomenys. Viena iš esamų problemų yra ta, kad kartais duomenys nėra pažymėti laike, todėl jų negalima naudoti klientų praradimo stebėjimui. Be to, reikalingas priėjimas prie klientų duomenų. Gauti duomenis apie klientus yra gana sudėtinga, nes daugelis įmonių baiminasi, kad tokie duomenys nepatektų konkurentams į rankas [14].

Skaitmeninių platformų tyrimas nors ir negali pateikti prognozuojamo modeliavimo, tačiau gali pateikti praradimą lemiančias priežastis. Savaime suprantama, kad klientai turi įvairių priežasčių, dėl kurių palieka įmonę. Didžiųjų duomenų nuspėjamieji modeliai gali padėti identifikuoti klientus, kurie gali atsisakyti teikiamų paslaugų, tačiau ne visi besitraukiantys klientai tai daro dėl tų pačių priežasčių ir su jais neturėtų būti elgiamasi vienodai. Todėl būtina turėti modelį, kuris leistų ne tik prognozuoti klientų praradimą, bet ir numatytų išlaikymo strategiją, atsižvelgiant į pasitraukimo veiksnius [14].

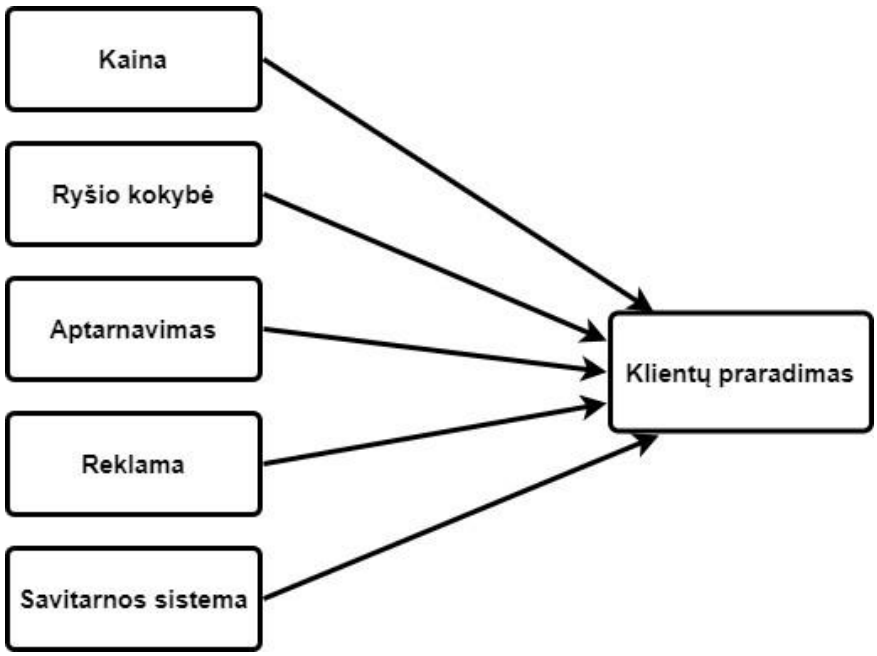
Išanalizavus mokslinę literatūrą, galima teigti, kad klientų praradimo valdymo modeliuose tyrinėjami įvairūs veiksniai, kurie lemia klientų netekimą. Verta paminėti, kad kliento apsisprendimą palikti įmonę dažniausiai lemia ne vienas, bet du ar net daugiau veiksnių. Veiksniai kai kuriais atvejais gali būti labai susiję, tai reiškia, kad išsprendus vieną problemą kita tampa nebeaktuali. Todėl būtina įvertinti ne tik individualių veiksnių įtaka klientų praradimui, bet ir jų tarpusavio ryšius.

1.4. Konceptualusis klientų praradimą telekomunikacijų paslaugų sektoriuje lemiančių veiksnių modelis

Remiantis literatūros apžvalga apie klientų praradimą, pasirinkti 5 veiksniai, turintys įtakos klientų praradimui telekomunikacijų paslaugų sektoriuje. Šie pasirinkimai paaiškinami taip:

1. **Kaina**. Šis veiksnys klientų išlaikymo kontekste analizuotas ne viename su telekomunikacijų paslaugomis susijusiame tyrime [15 – 16, 18]. Tyrimuose akcentuojama, kad klientai nėra linkę permokėti už paslaugas, o kainų didėjimas neigiamai atsiliepia lojalumui bei skatina ieškoti alternatyvų.
2. **Ryšio kokybė**. Ryšio kokybė, taip pat kaip ir kaina, yra dažnai tyrinėjamas veiksnys [15 – 16, 18]. Tai gana abstraktus veiksnys, kurį kai kurie tyrėjai labiau detalizuoja (pvz. interneto greitis, pokalbių kokybė). Šiuo atveju ryšio kokybės nuspręsta nedetalizuoti, o pasigilinti į bendra situaciją.
3. **Aptarnavimas**. Klientų patirtis telekomunikacijų srityje yra labai skirtinga, todėl aptarnaujant klientus telekomunikacijų įmonių darbuotojams būtinas prisitaikymas prie įvairių jų poreikių [15 – 16]. Tačiau ar tikrai tokia situacija, kai klientas gauna prastesnį aptarnavimą nei tikėjosi, gali lemti jo netekimą.
4. **Reklama**. Reklamos poveikis klientų praradimui analizuotas tik viename straipsnyje [16]. Dažnu atveju šis veiksnys nėra pagrindinis atsisakant teikiamų paslaugų, tačiau svarbu išsiaiškinti, ar jis yra aktualus ir atsiskleidžia nusivylusių klientų atsiliepimuose.
5. **Savitarnos sistema**. Daugelis klientų įprastas operacijas atlieka savitarnos svetainėje. Tai padeda sutaupyti laiko ir pastangų, tačiau sistemos nėra tobulos ir kartais gali sukelti net problemų [16]. Todėl šio veiksnio analizė gali padėti gauti naujų įžvalgų apie problemas, su kuriomis susiduria klientai naudodamiesi savitarnos sistema, bei ar tai gali paskatinti klientą pereiti pas konkurentus.

Atsižvelgiant į atliktą analizę, 1.6 paveiksle pateikiamas konceptualusis telekomunikacijų paslaugų klientų praradimą lemiančių veiksnių modelis.



1.6 pav. Konceptualusis klientų praradimą telekomunikacijų paslaugų sektoriuje lemiančių veiksnių modelis (sudaryta autoriaus)

Klientų praradimą lemiančių veiksnių identifikavimas leistų pasiūlyti jo mažinimą įgalinančius sprendimus. Klientams, besiskundžiantiems didele paslaugų kaina, būtų galima pasiūlyti individualius planus, geriausiai atitinkančius jų poreikius. Jei klientas visai nesinaudoja mobiliuoju internetu, tuomet jam reikėtų pasiūlyti labiau jo poreikius atitinkantį planą, kuris neturėtų didelio kiekio brangių mobiliųjų duomenų. Ryšio kokybės problemą verta spręsti identifikuojant pačias opiausias Lietuvos vietoves, kuriose klientai dažniausiai skundžiasi lėtu arba nestabiliu ryšiu. Identifikavus prasčiausias vietas, ieškoti techninių galimybių, kurios leistų pagerinti teikiamo ryšio kokybę. Siekiant gerinti aptarnavimo kokybę, reikėtų identifikuoti salonus bei juose dirbančius darbuotojus, kuriais klientai yra labiausiai nusivylę. Identifikuotiems salonams skirti didesnį dėmesį bei kelti juose dirbančių darbuotojų kompetenciją. Taip pat verta pagalvoti apie savitarnos svetainių tobulinimą. Dažniausiai skaitmeninėmis platformomis besinaudojantys žmonės turi pakankamas žinias naudotis savitarnos svetainėmis, todėl jų funkcionalumo didinimas leistų išvengti bereikalingo apsilankymo salonuose. Siekiant išvengti klientams nepatrauklios reklamos, būtina gilintis ne tik į nepasitenkinimą keliančius aspektus, bet ir į kanalus, kuriuose reklama buvo pastebėta. Gali būti, kad tos pačios reklamos rodymas televizijos transliacijoje sukels gerokai daugiau nepasitenkinimo nei vaizdavimas internetinėje erdvėje. Tai reiškia, kad neveiksmingos ir nepasitenkinimą keliančios reklamos problema gali slypėti ne pačioje reklamoje, bet tikslinės auditorijos pasirinkime.

Verta atkreipti dėmesį ne tik į individualius klientų praradimą lemiančius veiksniu, bet ir į jų tarpusavio ryšius. Klientų praradimą lemiančių veiksnių tarpusavio ryšių tyrime [16] pabrėžiama, kad kaina turi gana stiprų ryšį su kitais veiksniais, todėl tam tikrais atvejais jos didėjimas gali neturėti arba turėti labai mažai įtakos. Be to, kai kurie veiksniai yra gana abstraktūs, todėl esant galimybei juos patartina konkretizuoti. Vienas iš galimų pavyzdžių yra interneto greičio atskyrimas nuo ryšio kokybės, kuris gali padėti analizuojant klientus pagal skirtingas grupes [5]. Tam tikroms grupėms lėtas internetas gali sukelti rimtų problemų ar net paskatinti atsisakyti paslaugų, o kitos to gali net nepastebėti. Norint išsiaiškinti veiksnius, kurie lemia klientų nepasitenkinimą ir praradimą, pirmiausiai būtina analizuoti klientų atsiliepimus, suskirstant juos į teigiamus ir neigiamus. Tam tikslui naudojama sentimentų analizė, kuri aptariama kitame projekto skyriuje.

1.5. Sentimentų analizės panaudojimas tiriant klientų praradimą lemiančius veiksnius

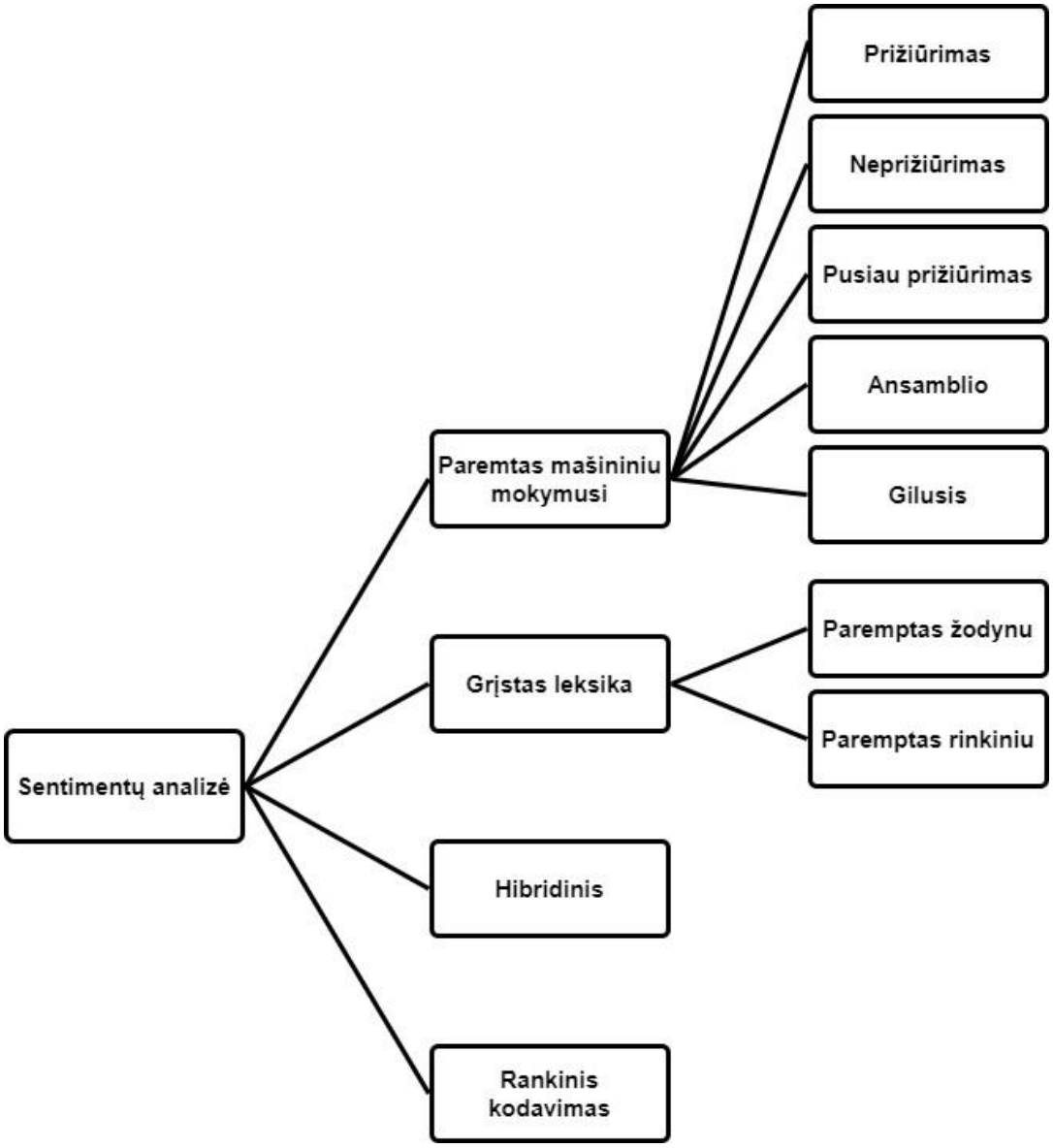
Sentimentų analizė apibūdinama kaip natūralios kalbos apdorojimo (angl. *NLP – natural language processing*) ir dirbtinio intelekto (angl. *AI – artificial intelligence*) forma, kuri padeda kompiuteriams interpretuoti ir suprasti žmonių kalbą. Konkrečiai kalbant apie socialinius tinklus, Lappeman‘as, Franco, Warner ir Sierra-Rubia [14] tyrime sentimentų analizei atlikti pasitelkė raktinius žodžius, kurie padėjo tekstui priskirti tam tikra sentimentą: teigiamą, neigiamą arba neutralų. Toks sentimentų identifikavimas naudojant raktinius žodžius yra vienas iš paprasčiausių ir tinkamas sudėtingesnės sintaksės kalboms. Kalbant apie anglų kalbą, dažnai moksliniuose straipsniuose įvardinami šie 2 metodai [20, 21]:

1. **Leksika grįstas sentimentų klasifikavimas naudojant žodžių ir emocijų asociaciją**. Šiam metodui gali būti naudojamas NRC emocijų leksikos (angl. *NRC emotion lexicon*) metodas, kuris turi sąrašą angliškų žodžių ir jų asociacijas su aštuoniomis emocijoms: džiaugsmu, pasitikėjimu, baime, nustebimu, liūdesiu, pasibjaurėjimu, pykčiu ir tikėjimu. Tai reiškia, kad sentimentas nustatomas ne pagal žodžius, bet pagal emocijas, su kuriomis analizuojamas žodis asocijuojasi.
2. **Leksika grįstas sentimentų klasifikavimas naudojant poliškumą**. Šis metodas taip pat yra grįstas žodžiais, tačiau papildomai remiasi leksika arba žodynu, kuriame yra nurodytas žodžių

poliškumas. Tai reiškia, kad pozityvūs išsireiškimai yra traktuojami kaip skaičiai nuo 0 iki 1, o negatyvūs nuo -1 iki 0. Galiausiai, sudėjus visas reikšmes nustatoma ar, tekstas turi daugiau teigiamos ar neigiamos emocijos.

Žinoma, toks kalbos apdorojimo būdas nesugeba iki galo suprasti žmogaus pokalbio niuansų, todėl dažnai prireikia rankinio testavimo ir klaidų taisymo. Dažniausiai problemos kyla analizuojant tekstus, kuriuose pasitaiko humoras ar sarkazmas. Šioms problemoms spręsti dažnai pasirenkama atlikti rankinį sentimentų priskyrimą. Tai padeda išspręsti anksčiau minėtų metodų problemas, tačiau susiduriama su kitomis: ilgai trunkantis procesas, netinka turint didelius duomenų kiekius, galimos žmogiškosios klaidos ir skirtingi vertintojų požiūriai [22].

Kitame analizuotame straipsnyje apie sentimentų analizę ir nuomonės gavybą viešojo saugumo srityje [23] išskiriamos 4 didelės kategorijos, į kurias skirstomi sentimentų analizės metodai. Išanalizavus tekste minimas sentimentų analizės grupes, sudarytas vizualus modelis 1.7 paveiksle, kuriame pateikti pagrindiniai sentimentų analizės būdai.



1.7 pav. Sentimentų analizės metodai (Suhaimin, Hijazi ir kt. [23])

1. **Sentimentų analizės metodas paremtas mašininiu mokymusi** (angl. *Machine learning-based*)

    1.1. **Prižiūrimas mokymasis** (angl. *Supervised learning*). Prižiūrimas mašininis mokymasis naudoja jau paruoštą duomenų rinkinį, kuriame kiekvienas tekstas turi priskirtą sentimentą. Toks tekstas yra paduodamas mašininiam modeliui ir modelis apmokomas klasifikuoti panašaus tipo tekstus.

    1.2. **Neprižiūrimas mokymasis** (angl. *Unsupervised learning*). Neprižiūrimas mokymosi metodas nereikalauja treniravimo duomenų imties su jau priskirtais sentimentais. Jis naudoja paslėptas struktūras arba semantines asociacijas nepažymėtuose duomenyse ir gali būti pritaikytas tekstiniams duomenims be rankinio įsikišimo.

    1.3. **Pusiau prižiūrimas mokymasis** (angl. *Semi-supervised learning*). Iš dalies prižiūrimas mokymasis naudoja nedidelį kiekį jau pažymėtų arba anotuotų duomenų ir didesnį nepažymėtų duomenų kiekį klasifikavimui.

    1.4. **Ansamblio mokymasis** (angl. *Ensemble learning*). Ansamblio metodo idėja yra sujungti kelių bazinių modelių rezultatus, kad būtų sukurtas vienas rezultatas, kuris duoda geriausią tikslumą.

    1.5. **Gilusis mokymasis** (angl. *Deep learning*). Gilus mokymasis yra mašininio mokymosi pogrupis, kuriame naudojami gilieji neuroniniai tinklai.

2. **Sentimentų analizės metodas, grįstas leksika** (angl. *Lexicon approaches*)

    2.1. **Paremtas žodynu** (angl. *Dictionary-based*). Žodynu paremti leksikos metodai naudoja žodyną, kuris integruoja terminų poliškumą. Kai tekste aptinkamas žodis, atliekama paieška žodyne ir tada apskaičiuojamas nuotaikos balas.

    2.2. **Paremtas rinkiniu** (angl. *Corpus-based*). Šis metodas remiasi dideliu kalbos tekstynu. Šis tekstynas gali apimti įvairius kalbos duomenis, pavyzdžiui, knygas, straipsnius, interneto tinklalapius ar kitą kalbos turinį. Šių teksto duomenų analizė padeda modeliui išmokti kalbos struktūrą, taisykles ir sąryšius, o tai leidžia jam geriau suprasti įvairų tekstą.

3. **Hibridinis sentimentų analizės metodas** (angl. *Hybrid*). Remiantis literatūra, kai kuriuose aprašytuose darbuose buvo naudojamas hibridinis būdas, kuris apjungia mašininį mokymąsi ir leksika grįstus metodus. Šis metodas buvo sukurtas siekiant kompensuoti mašininio mokymosi ir leksika grįstų metodų trūkumus, kai jie naudojami atskirai. Tai yra dažniausiai naudojamas metodas po prižiūrimo mokymosi.

4. **Sentimentų analizės metodas paremtas rankiniu kodavimu** (angl. *Manual coding*). Tai procesas, kurio metu žmonės rankiniu būdu priskiria kodus, žymes arba kategorijas duomenims.

Nors didžioji dalis analizuotų straipsnių koncentruojasi tik į anglų kalbą, tačiau galima rasti tyrimų, kuriuose gilinamasi ir į lietuvių kalbą. Štrimaičio, Stefanovič ir kt. [27] straipsnyje pabrėžiama, kad lietuvių kalba yra labai sudėtinga dėl žodžių formų įvairovės ir galimų įvairių sakinio sandarų. Dėl šių priežasčių lietuvių kalbos tekstų analizė vis dar yra gana prastai išvystyta. Be to, pabrėžiama, kad lietuvių kalbai atlikti sentimentų analizę yra gerokai sudėtingiau, o pasiektas tikslumas labai tikėtina bus mažesnis negu panašiuose eksperimentuose, kuriuose naudojama anglų kalba [28]. Be jau minėtų ypatybių, lietuvių kalba dar pasižymi: mažybiniais žodžiais (pvz. dukra gali būti pavadinta

dukružėle), prasmės skirtumais dėl diakritinių ženklų (pvz. panaudojus diakritinį ženklą gaunamas žodis karštas, o be jo karstas), įvairiomis galūnėmis ir priešdėliais. Kitame skyriuje aptariamas mašininis mokymasis bei modeliai, kurie dažniausiai naudojami sentimentų analizės tyrimuose.

1.6. Mašininis mokymasis klientų atsiliepimų klasifikavime

Galima rasti nemažai mokslinių straipsnių, kuriuose gilinamasi į sentimentų analizę, ir taip pat išbandomi bent keli mašininio mokymosi metodai bei įvertinamas jų tikslumas. Kadangi mašininio mokymosi modeliai nesugeba suprasti tekstinės informacijos, todėl būtina atlikti vektorizavimą. Šiai užduočiai įgyvendinti galima rinktis iš daugybės galimų metodų. Analizuotoje literatūroje dažniausiai naudojami šie:

1. **Žodžių krepšelis** (angl. *Bag of words, BOW*). Tai yra teksto reprezentavimo metodas, kuris konvertuoja tekstą į žodžių dažnio vektorių be jokios eilės informacijos [22, 33].
2. **Termino dažnis – atvirkštinis dokumento dažnis** (angl. *Term frequency-inverse document frequency, TF-IDF*). Tai yra statistinė teksto analizės technika, naudojama įvertinti žodžių svarbą dokumente, atsižvelgiant į jų pasikartojimus visuose dokumentuose [22, 25 – 26, 29 – 31].
3. **Word2Vec**. Populiari mašininio mokymo technika, skirta įveikti žodžių reprezentavimo iššūkius. Ši technika yra skirta konvertuoti žodžius į vektorius, taip leisdama kompiuteriui suprasti žodžių semantiką pagal jų kontekstą. Pagrindinė idėja yra ta, kad žodžiai, kurie dažnai pasirodo arti vienas kito, turi panašius prasmės aspektus [33, 37].
4. **FastText**. Klasifikavimo ir vektorizavimo algoritmas, sukurtas „Facebook“ mokslininkų. Šis metodas yra paremtas ne tik žodžių, bet ir jų subkomponentų n-gramų analize [28, 33].
5. **GloVe**. Pagrindinė šio metodo idėja yra sukurti vektorius, atspindinčius žodžių semantinį panašumą ir santykius. Jis analizuoja, kaip dažnai dvi žodžių poros sutinkamos kartu tekste ir kiek artimos jos yra viena kitai [40]

Kalbant apie klasifikavimo modelius, literatūroje jų galima rasti daug ir išsirinkti iš tokios didelės gausos yra gana sunku. Išanalizavus ne vieną mokslinį straipsnį, sudaryta 1.1 lentelė, kurioje pateikiamas literatūros šaltinis, naudotas duomenų rinkinys, jo dydis bei klasifikavimo metodai.

1.1 lentelė. Analizuotoje literatūroje naudojami klasifikavimo metodai

Šaltinis	**Duomenų šaltinis**	**Duomenų rinkinio dydis**	**Metodai**
Sidya, Fanany, Budi (2015) [24]	Twitter	10 004	Bajeso metodas (angl. *Naïve Bayes*) Atraminių vektorių klasifikatorius (angl. *Support vector machine*) Sprendimų medis (angl. *Decision tree*)
Qamar, Sohali (2017) [25]	Twitter	1 331	Bajeso metodas (angl. *Naïve Bayes*) Logistinė regresija (angl. *Logistic regression*) K-artimiausių kaimynų (angl. *K-nearest neighbor*)
Kapočiūtė-Dzikienė, Damaševičius, Woźniak (2019) [33]	lrytas	10 570	Bajeso metodas (angl. *Naïve Bayes*) Atraminių vektorių klasifikatorius (angl. *Support vector machine*) CNN (angl. *Convolutional neural network*)

			LSTM (angl. *Long short term memory*)
Dzisevič, Šešok (2019) [30]	Alio, Skelbiu	10 000	Neuroniniai tinklai (angl. *Neural network*)
Abou el Kassem, Ali Hussein, Mostafa Abdelrahman, Kamal Alsheref (2020) [34]	Facebook	352	Bajeso metodas (angl. *Naïve Bayes*) Logistinė regresija (angl. *Logistic regression*) Gilusis mokymasis (angl. *Deep learning*)
Jeelall, Cheerkoot-Jalim (2020) [35]	Twitter, MedHelp	3 200	Bajeso metodas (angl. *Naïve Bayes*) Atsitiktiniai miškai (angl. *Random forest*) Atraminių vektorių klasifikatorius (angl. *Support vector machine*) K-artimiausių kaimynų (angl. *K-nearest neighbor*)
Štrimaitis, Stefanovič, Ramanauskaitė, Slotkienė (2021) [27]	Lietuviški naujienų portalai	10 375	Bajeso metodas (angl. *Naïve Bayes*) Atraminių vektorių klasifikatorius (angl. *Support vector machine*) LSTM (angl. *Long short term memory*)
Umarania, Juliana, Deepa, (2021) [31]	Kaggle	1 000	Bajeso metodas (angl. *Naïve Bayes*) Logistinė regresija (angl. *Logistic regression*) Atsitiktiniai miškai (angl. *Random forest*) Atraminių vektorių klasifikatorius (angl. *Support vector machine*) K-artimiausių kaimynų (angl. *K-nearest neighbor*) Sprendimų medis (angl. *Decision tree*) CNN (angl. *Convolutional neural network*) LSTM (angl. *Long short term memory*)
Aljedaani, Rustam, Mkaouer, Mkaouer, Ghallab, Rupapara, Washington, Lee, Ashraf (2022) [22]	Kaggle (Twitter)	14 640	Sprendimų medis (angl. *Decision tree*) Atsitiktiniai miškai (angl. *Random forest*) Papildomų medžių klasifikatorius (angl. *Extra trees classifier*) Gradiento didinimo klasifikatorius (angl. *Gradient boosting classifier*) Logistinė regresija (angl. *Logistic regression*) Atraminių vektorių klasifikatorius (angl. *Support vector machine*)
Biradar, Gorabal, Gupta (2022) [26]	Twitter	-	Bajeso metodas (angl. *Naïve Bayes*) Sprendimų medis (angl. *Decision tree*) OneR (angl. *One rule*)
Kapočiūtė-Dzikienė, Salimbajevs (2022) [28]	Įvairūs lietuviški portalai	18 981	FFNN (angl. Feed forward neural network) CNN (angl. *Convolutional neural network*)
Diekson, Bagas Prakoso, Qalby Putra, Al Fadel Syaputra, Achmad, Sutoyo (2022) [29]	Twitter	1 200	Atraminių vektorių klasifikatorius (angl. *Support vector machine*) Bajeso metodas (angl. *Naïve Bayes*) Logistinė regresija (angl. *Logistic regression*)
Mahmud, Islam, Jaman Bonny, Khatun Shorna, Hossain Omi, Rahman (2022) [40]	Facebook, forumai, žiniasklaida	2 272	CNN (angl. *Convolutional neural network*) LSTM (angl. *Long short term memory*)
AminiMotlagh, Shahhoseini, Fatehi (2023) [32]	Twitter	6 980	K-artimiausių kaimynų (angl. *K-nearest neighbor*) Sprendimų medis (angl. *Decision tree*)

			Atraminių vektorių klasifikatorius (angl. *Support vector machine*) Bajeso metodas (angl. *Naïve Bayes*) Metodų ansamblis (angl. *Ensamble methods*)
Bikku, Jarugula, Kongala, Tummala, Donthiboina (2023) [36]	Twitter	1.6 mln.	BERT (angl. *Bidirectional encoder representations from transformers*) Logistinė regresija (angl. *Logistic regression*) Atsitiktiniai miškai (angl. *Random forest*) Bajeso metodas (angl. *Naïve Bayes*) Atraminių vektorių klasifikatorius (angl. *Support vector machine*)
Lubis, Fatmi, Witarsyah (2023) [37]	Twitter	12 900	BERT (angl. *Bidirectional encoder representations from transformers*) LSTM (angl. *Long short term memory*) CNN (angl. *Convolutional neural network*) Logistinė regresija (angl. *Logistic regression*) Bajeso metodas (angl. *Naïve Bayes*) Atsitiktiniai miškai (angl. *Random forest*)
Singh, Srivastava, Aman, Dubey (2023) [38]	Developers Bay	500	BERT (angl. *Bidirectional encoder representations from transformers*)
Bhattacharjee, Paul, Kumar (2023) [39]	Twitter, Reddit, YouTube	39 565	CNN (angl. *Convolutional neural network*) LSTM (angl. *Long short term memory*)
Taherkhani, Daneshvar, Khalili, Sanaei (2023) [41]	Socialiniai tinklai, viešbučių interneto svetainės	6 000	Atsitiktiniai miškai (angl. *Random forest*) Gradiento didinimo klasifikatorius (angl. *Gradient boosting classifier*) Bajeso metodas (angl. *Naïve Bayes*) Sprendimų medis (angl. *Decision tree*) K-artimiausių kaimynų (angl. *K-nearest neighbor*)
Alshamari (2023) [42]	Twitter	20 000	BiLSTM (angl. *Bidirectional long short term memory*) CNN (angl. *Convolutional neural network*) LSTM (angl. *Long short term memory*) GRU (angl. *Gated recurrent unit*)

Sudarytoje mokslinių straipsnių lentelėje galima pastebėti, kad dažniausiai pasirenkamas duomenų šaltinis – „Twitter“ socialinis tinklas. Šis socialinis tinklas siūlo patogią API sąsają, kuri ir lemia tokį didelį šio socialinio tinklo populiarumą analizuotoje literatūroje. Be to, verta atkreipti dėmesį į giliojo mokymosi (angl. *Deep learning*) panaudojimą tekto analizėje. Nors klasikiniai mašininio mokymosi modeliai vis dar dominuoja, tačiau vis daugiau dėmesio kreipiama į gilųjį mokymąsi bei išbandomas jo veiksmingumas praktiniais tyrimais. Ypač džiugu tai, kad tarp analizuotų straipsnių pavyko rasti net keletą, kuriuose buvo tiriama lietuvių kalba pasitelkiant gilųjį mokymąsi [27 – 28, 33]. Paskutiniame literatūros analizės skyriuje trumpai aptariami neprižiūrimo mašininio mokymosi metodai, kurie naudojami temos modeliavime.

1.7. Temos modeliavimas analizuojant neigiamus klientų atsiliepimus

Siekiant išsiaiškinti klientų atsiliepimuose dominuojančias temas dažnai naudojami temos modeliavimo metodai. Wayasti, Surjandari ir Zulkamain‘as [43] tyrime panaudojo tekstinius duomenis iš „Twitter“ socialiniame tinkle paskelbtų žinučių, skirtų vienam iš pavėžėjimo paslaugų

teikėjų Indonezijoje. Modeliavimui pasirinktas atviras Dirichlet pasiskirstymo (angl. *Latent Dirichlet Allocation, LDA*) metodas, kuris padėjo identifikuoti net devynias atsiliepimuose dominuojančias temas. Šio tyrimo rezultatai įmonei leido sužinoti, kokia informacija apie juos skelbiama socialinėje erdvėje, ir ja remiantis pagerinti teikiamų paslaugų kokybę.

Yamunathangam, Bharathi Priya, Shobana ir Latha [44] suskirstė įvairius temos modeliavimo metodus į keturias grupes. Toliau pateiktoje 1.2 lentelėje išvardijamos visos grupės bei kiekvienai grupei priklausantys modeliavimo metodai.

1.2 lentelė. Įvairių temos modeliavimo metodų klasifikavimas

Grupė	**Kategorija**	**Modeliavimo metodas**
Standartiniai temos modeliai	Netikimybiniai	VSM (angl. *Vector space model*) LSI (angl. *Latent semantic indexing*) NMF (angl. *Non-negative matrix factorization*)
	Tikimybiniai	PLSA (angl. *Probabilistic latent semantic analysis*) LDA (angl. *Latent dirichlet allocation*) MM (angl. *Multinomial mixture*)
Klasterizavimu paremti temos modeliai	Netikimybiniai	TermCut WordCom
	Tikimybiniai	GSDMM (angl. *Gibbs sampling dirichlet multinomial mixture*) GPU – DMM (angl. *Generalized Polya Urn -dirichlet multinomial mixture*) BTM (angl. *Biterm topic model*) PYPM (angl. *Pitman-Yor orocess mixture*)
Savaime agreguojantys temos modeliai	Tikimybiniai	SATM (angl. *Self-aggregation based topic model*) PTM (angl. P*seudo-document-based topic model*)
Giliuoju mokymusi grįsti temos modeliai	Tikimybiniai	RNN + BTM (angl. *Recurrent neural network + biterm topic model*)
	Tikimybiniai ir netikimybiniai	LTMF (angl. *Long short-term memory topic matrix factorization*)

Tokį didelį temos modeliavimo metodų pasirinkimą lemia tai, kad juos galima labai plačiai panaudoti. Šie metodai naudojami: rekomendacinėse sistemose, sentimentų analizėje, nepageidaujamų laiškų aptikime ar net ieškant genomo sekos panašumų [44].

Išanalizavus mokslinius straipsnius, galima teigti, kad klientų praradimo tema yra sparčiai populiarėjanti ir tikėtina, kad ateityje jos paklausa vis labiau augs. Dalis tyrėjų pasirenka vidinius įmonių duomenis, kuriuose galima rasti labai detalią informacija apie klientus, tačiau tokius duomenis gauti yra didelis iššūkis. Kita dalis tyrėjų nusprendžia rinkti informaciją internete, kuri yra prieinama plačiajai visuomenei. Pastebėta, kad pastaruoju metu vis dažniau pasirenkama analizuoti viešai prieinamą informaciją ir tai leidžia gauti gerokai daugiau informacijos apie klientų požiūrį į įmonę ir

jos teikiamas paslaugas. Tačiau nusprendus analizuoti viešai prieinamą informaciją susiduriama su iššūkiais: komplikuotas duomenų rinkimas, sentimentų priskyrimo netikslumai, platus teksto vektorizavimo bei klasifikavimo metodų pasirinkimas. Šios problemos skatina tyrėjus gilintis į natūralios kalbos apdorojimą ir ieškoti metodų bei technikų, kurios leistų gauti kuo tikslesnius rezultatus. Šiame projekte nuspręsta taikyti:

- Duomenų rinkimą iš Lietuviškų interneto svetainių bei „Facebook“ socialinio tinklo;
- Sentimentų analizę naudojant rankinį kodavimą;
- Atsiliepimų vektorizavimą naudojant 4 metodus: žodžių krepšelį, termino dažnį – atvirkštinį dokumento dažnį, Word2Vec (skip-gram) ir Word2Vec (CBOW);
- Klasifikavimą naudojant 5 skirtingus modelius: gradiento didinimą, XGBoost, atsitiktinius miškus, logistinę regresiją ir k – artimiausių kaimynų;
- Atsiliepimų konteksto analizę naudojant Word2Vec (skip-gram) techniką;
- Temos modeliavimą naudojant LDA metodą.

2. Tyrimo metodologija

Šioje projekto dalyje pateikiami naudoti duomenys bei detaliau aprašomos technologijos, metodai ir modeliai, kuriuos pasirinkta išbandyti klientų praradimą lemiančių veiksnių analizėje.

2.1. Duomenų rinkimas, apdorojimas bei sentimentų analizė

Tyrime nuspręsta rinkti duomenis iš keliolikos interneto svetainių, kuriose klientai palieka daugiausiai atsiliepimų apie didžiausias Lietuvoje telekomunikacijų įmones bei jų teikiamas paslaugas. Duomenų rinkimui pasirinkta Python programavimo kalba bei keletas bibliotekų, kurios padėjo gauti atsiliepimus iš interneto svetainių. Duomenų rinkimui iš „Facebook“ panaudota *facebook-scraper*, o visoms kitoms svetainėms – *BeautifulSoup* biblioteka.

Pirmas žingsnis, surinkus duomenis, yra jų išvalymas. Seniausi rasti atsiliepimai parašyti 2010 metais, todėl nuspręsta sumažinti duomenų imtį ir analizuoti tik tuos atsiliepimus, kurie parašyti ne anksčiau negu 2014 metais. Originalus tekstas turi daug nereikšmingos informacijos, kuri apsunkina analizę, todėl būtina pašalinti tai, kas nesukuria jokios pridėtinės vertės. Nuspręsta atlikti tokius teksto apdorojimus:

- Pašalinti nuorodas į kitas interneto svetaines;
- Pašalinti skyrybos ir specialius simbolius (pvz. *, %, $);
- Pašalinti skaičius;
- Pašalinti nereikšmingus žodžius (pvz. tas, ir, ar);
- Paversti tekstą į mažąsias raides;
- Pašalinti per didelius tarpus tarp žodžių;
- Atlikti teksto lemavimą (angl. *lemmatization*), kuris įvairių gramatinių formų žodžius paverčia į pagrindinę (pavyzdžiui žodis „atsiliepime“ paverčiamas į „atsiliepimas“);
- Pašalinti žodžius, kurie visame duomenų rinkinyje pasikartoja mažiau nei 10 kartų.

Sutvarkius tekstą galima pradėti sentimentų analizės žingsnį. Sentimentams dažniausiai naudojamos 2 klasės: neigiama ir teigiama. Kai kur dar išskiriama trečioji, kuri reiškia neutralią emociją. Literatūroje galima rasti daugybę priemonių, kurios automatiškai sugeba atpažinti tekste dominuojančią emociją ir tai atlieka gana tiksliai. Deja, bet tokios priemonės kol kas geriausiai veikia tik su anglų kalba, o daugelio kitų kalbų iš vis nesugeba apdoroti arba tai atlieka labai prastai. Šiai problemai spręsti dažnai pasirenkami žodynai, kurie pagal tam tikrus raktinius žodžius ar frazes sugeba nustatyti teksto sentimentą. Nors šis būdas nėra pats tiksliausias, tačiau žodyną galima susidaryti pačiam ir pasiekti tikrai gana gerus sentimentų analizės rezultatus. Taip pat naudojamas ir rankinis kodavimas, kurio metu atsakingi asmenys priskiria visiems komentarams atitinkamus sentimentus. Šis būdas yra bene tiksliausias, tačiau susiduriama su keliomis problemomis: ilgai trunkantis procesas ir skirtingas vertintojų požiūris.

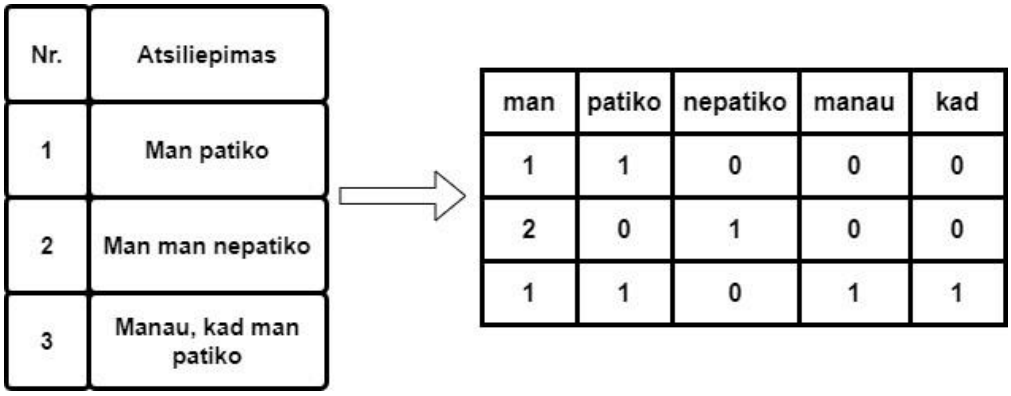
Kadangi šiame projekte analizuojami klientų atsiliepimai yra tik lietuvių kalba, todėl automatinių priemonių sentimentams priskirti nepavyko rasti. Dėl šios priežasties sentimentus nuspręsta priskirti rankiniu kodavimu analizuojant kiekvieną atsiliepimą atskirai. Tiems klientų atsiliepimams, kuriems nepavyko priskirti nei neigiamo, nei teigiamo sentimento, įvardinti kaip neutralūs ir jų tolimesniuose projekto etapuose nuspręsta nenaudoti.

2.2. Duomenų vektorizavimo metodai

Norint pasiekti gerus klasifikavimo rezultatus, būtina ne tik išbandyti skirtingus klasifikavimo modelius, bet ir atkreipti dėmesį į duomenų vektorizavimo arba kitaip vadinamus požymių išskyrimo metodus. Šiame projekte nuspręsta išbandyti toliau aprašytus metodus ir tuomet palyginti, kuris požymių išskyrimo metodas buvo tinkamiausias.

2.2.1. Žodžių krepšelis

Žodžių krepšelis (angl. *Bag of words, BOW*) tai natūralios kalbos apdorojimo ir teksto analizės metodas, kuris susideda iš žodžių arba frazių rinkinio, vadinamojo žodžių krepšelio, kuris atspindi atsiliepimo turinį. Šis metodas ignoruoja teksto struktūrą ir kontekstą, bet susikoncentruoja ties žodžių pasikartojimu ir jų dažnumu atsiliepime (2.1 pav.).



2.1 pav. Žodžių krepšelio metodo logika

Pagrindinė šio metodo idėja yra ta, kad kiekvienas atsiliepimas gali būti apibūdinamas kaip žodžių dažnių rinkinys, ir atsiliepimų palyginimas gali būti atliekamas remiantis šių dažnių panašumu.

2.2.2. Termino dažnis – atvirkštinis dokumento dažnis

Anksčiau aptartas žodžių krepšelio metodas yra paprastas, tačiau jis turi gana didelį trūkumą, nes traktuoja visus žodžius vienodai. Dėl to negalima atskirti labai dažnų ir retų žodžių atsiliepime. Skirtingai nei žodžių krepšelio metodas, termino dažnio – atvirkštinio dokumento dažnio metodas (angl. *Term Frequency Inverse Document Frequency, TF-IDF*) atsižvelgia į kiekvieno žodžio svarbą atsiliepime. Terminas TF reiškia termino dažnį, o IDF – atvirkštinį dokumento dažnį.

Norint suprasti TF-IDF, pirmiausiai reikia atskirai aptarti du terminus:

- Termino dažnis (TF)
- Atvirkštinis dokumento dažnis (IDF)

Termino dažnis (TF) įvertina, kiek kartų konkretus žodis pasikartoja atsiliepime. Dažnai tai yra tiesiog žodžių pasikartojimų skaičius, padalintas iš bendro žodžių skaičiaus dokumente:

$$TF = \frac{t}{d}, \tag{1}$$

čia t – kiek kartų žodis randamas atsiliepime, d – žodžių skaičius atsiliepime.

Atvirkštinis dokumento dažnis (IDF) įvertina, kaip dažnai žodis randamas visuose atsiliepimuose. Ši reikšmė yra skaičiuojama kaip bendras atsiliepimų skaičius, padalintas iš žodžio pasikartojimų skaičiaus visuose atsiliepimuose:

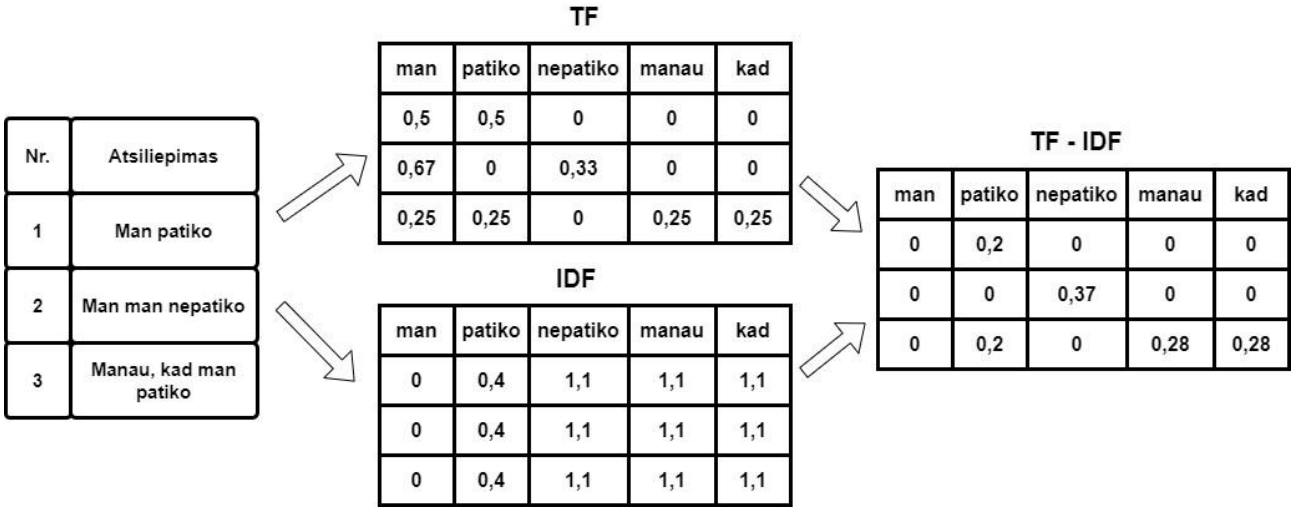
$$IDF = \log\left(\frac{n}{a}\right), \tag{2}$$

čia n – bendras atsiliepimų skaičius, a – atsiliepimų, turinčių analizuojamą žodį, skaičius.

TF-IDF įvertis dokumente apskaičiuojamas sudauginus TF ir IDF reikšmes (3 formulė). Tai leidžia nustatyti žodžius, kurie pasikartoja dažnai dokumente, bet retai visame duomenų rinkinyje.

$$TFIDF = TF \times IDF, \tag{3}$$

Pateikiamas praktinis pavyzdys 2.2 paveiksle, kai pasirinktiems keliems vartotojų atsiliepimams pritaikomas TF-IDF metodas.



2.2 pav. TF-IDF metodo logika

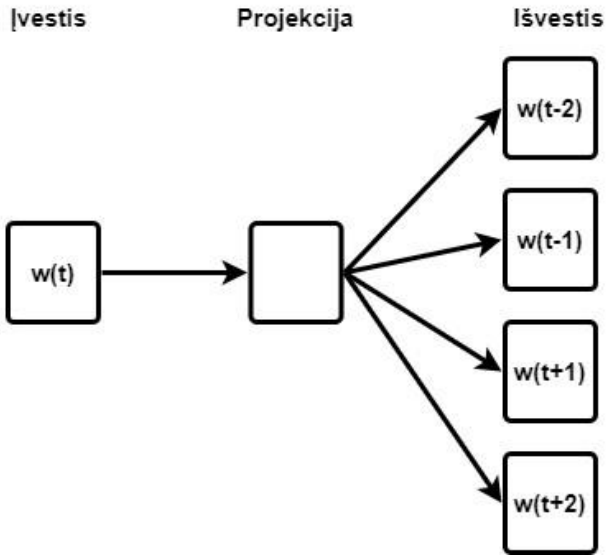
Kaip galima pastebėti pavyzdyje, TF-IDF suteikia didesnį svorį žodžiams, kurie dažnai pasitaiko konkrečiame atsiliepime, bet retai pasitaiko kituose atsiliepimuose. Tai padeda atskirti svarbius žodžius nuo dažnai pasitaikančių žodžių, gerinant teksto analizės rezultatus.

2.2.3. Word2Vec

Word2Vec yra natūralios kalbos apdorojimo technologija, skirta žodžių reprezentacijai ir semantinio panašumo modeliavimui. Technologija sukurta 2013 metais „Google“ mokslininkų. Šis metodas žodžius, kurie pasirodo tuose pačiuose kontekstuose, identifikuoja kaip turinčius panašią semantiką ir prasmę.

Word2Vec modelio sukūrimui yra naudojami du pagrindiniai būdai:

1. **Skip-gram.** Šis algoritmas bando prognozuoti aplinkinius žodžius pagal konkretų jam pateiktą žodį (2.3 pav.). Tai padeda išsiaiškinti, koks kontekstas dažniausiai minimas šalia analizuojamo žodžio.



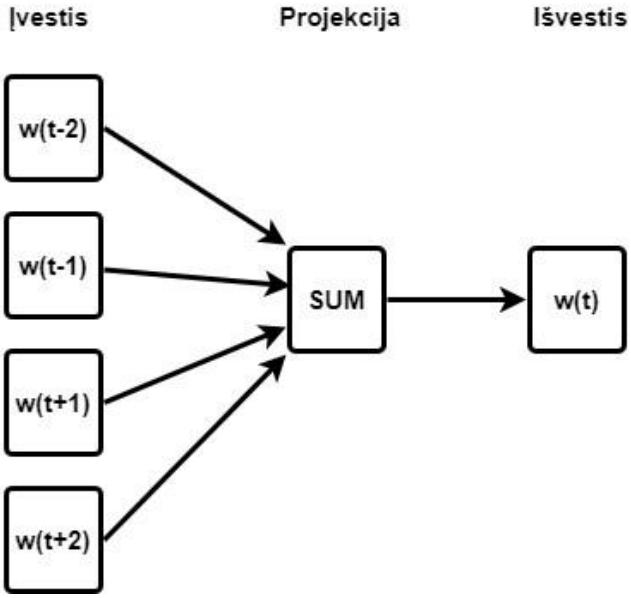
2.3 pav. Skip-gram architektūra

Toliau pateiktoje 2.1 lentelėje trumpai aprašomi skip-gram architektūros sluoksniai.

2.1 lentelė. Skip-gram architektūros sluoksniai

Sluoksnis	**Sluoksnio paskirtis**	**Modelio tikslas**
Įvestis	Tikslinio žodžio vektoriaus radimas.	Koreguoti vektorius taip, kad tikimybė teisingai nuspėti konteksto žodžius būtų kuo didesnė.
Projekcija (paslėptas sluoksnis)	Įvesties vektorius projektavimas į paslėptą sluoksnį.	
Išvestis	Kiekvienai kontekstinio lango pozicijai (w(t-2), w(t-1)...) išvesti vektorių, kuris geriausiai reprezentuoja tikslinį žodį.	

2. **Tęstinis žodžių krepšelis** (angl. *Continuous bag of words, CBOW*), priešingai nei skip-gram, CBOW bando prognozuoti konkretų žodį pagal aplinkinius žodžius (2.4 pav.). Tai reiškia, kad remiantis aplinkiniais žodžiais, modelis bando atkurti konkretų žodį.



2.4 pav. CBOW architektūra

Toliau pateiktoje 2.2 lentelėje trumpai aprašomi CBOW architektūros sluoksniai.

2.2 lentelė. CBOW architektūros sluoksniai

Sluoksnis	**Sluoksnio paskirtis**	**Modelio tikslas**
Įvestis	Susieti visus konteksto žodžius su unikaliu vektoriumi (paprastai inicijuojamu atsitiktinai mokymo pradžioje).	Koreguoti vektorius taip, kad sujungtas konteksto vektorius būtų artimas tikrojo tikslinio žodžio vektoriui.
Projekcija (paslėptas sluoksnis)	Sujungti konteksto žodžių vektorius, kad būtų sukurtas naujas vektorius. Dažniausias derinimo būdas yra vidurkio paėmimas, tačiau juos galima ir sumuoti.	
Išvestis	Sujungtas vektorius naudojamas nuspėti tikslinį žodį w(t). Tai vyksta lyginant gautą išvesties vektorių su visais žodyno vektoriais ir vertinant, kuris yra artimiausias.	

Trumpai apibendrinant word2vec modelį, kiekvienam žodžiui šis modelis priskiria vektorinę reprezentaciją, kurioje žodis yra atstovaujamas kaip skaičių vektorius. Šie vektoriai yra pasirinkti taip, kad panašūs žodžiai turėtų panašius vektorius, o nepanašūs – skirtingus. Kadangi visi vektorizavimo metodai apžvelgti, galima pareiti prie klasifikavimo modelių.

2.3. Klasikiniai klasifikavimo modeliai

Klasifikavimas (angl. *Classification*) – tai procesas, kurio tikslas yra priskirti kategoriją arba klasę naujiems stebėjimams arba duomenims, remiantis jau turimais suklasifikuotais duomenimis. Toliau pateikiami visi šiame projekte naudoti klasifikavimo metodai bei aprašomas jų veikimo principas bei subtilybės.

2.3.1. Gradiento didinimas

Gradiento didinimas (angl. *Gradient boosting*) yra mašininio mokymosi technika, dažniausiai naudojama klasifikavimo ir regresijos užduotims spręsti. Modelis sudaromas iš silpnesnių prognozavimo modelių – sprendimų medžių rinkinių. Pagrindinė gradiento didinimo idėja yra sukurti modelį, kuris pašalintų ankstesnių modelių klaidas.

Toliau pateiktuose žingsniuose detaliau aprašomas šio modelio veikimas:

1. Pirmasis žingsnis apima bazinio modelio kūrimą duomenų rinkiniui prognozuoti. Kad būtų lengviau atlikti skaičiavimus, imamas tikslinio stulpelio vidurkis ir manoma, kad tai yra numatoma vertė. Pirmo žingsnio matematinė interpretacija:

$$F_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, \gamma), \tag{3}$$

   čia L – nuostolio funkcija, γ (gama) – numatoma vertė, argmin – prognozuojama rasti vertė arba γ, kurios nuostolių funkcija yra minimali.

Kadangi tikslinis stulpelis yra tęstinis, nuostolių funkcija apskaičiuojama pagal formulę:

$$L = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 . \quad (4)$$

2. Sukamas ciklas kai *m*=1 iki *M* reikšmės:
   2.1. Randamas likutis (angl. *Pseudo-residuals*):

$$r_{im} = -\left[\frac{\partial L\big(y_i, F(x_i)\big)}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} . \quad (5)$$

   2.2. Šis žingsnis apima kiekvieno sprendimų medžio lapo išvesties verčių radimą. Svarbu rasti visų lapų išvestį, nes vienas lapas gali turėti daugiau nei vieną likutį. Nesvarbu, ar yra tik vienas skaičius ar daugiau, rezultatą galime lengvai apskaičiuoti imdami visų lapo reikšmių vidurkį.

$$\gamma_m = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L\big(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)\big), \quad (6)$$

   čia $\gamma h_m(x_i)$ – sprendimų medis, susidaręs per likutį, m – sprendimų medžio numeris.
   2.3. Paskutiniame etape atnaujinamos ankstesnio modelio prognozės. Jei nepasiekta *M* reikšmė, tuomet grįžtama į 2.1 žingsnį. Kitu atveju pateikiama prognozė:

$$F_m(x) = F_{m-1}(x) + \gamma_x h_m(x), \quad (7)$$

   čia m – sprendimų medžių skaičius, $F_m(x) = F_{m-1}(x)$ – bazinio modelio prognozės.

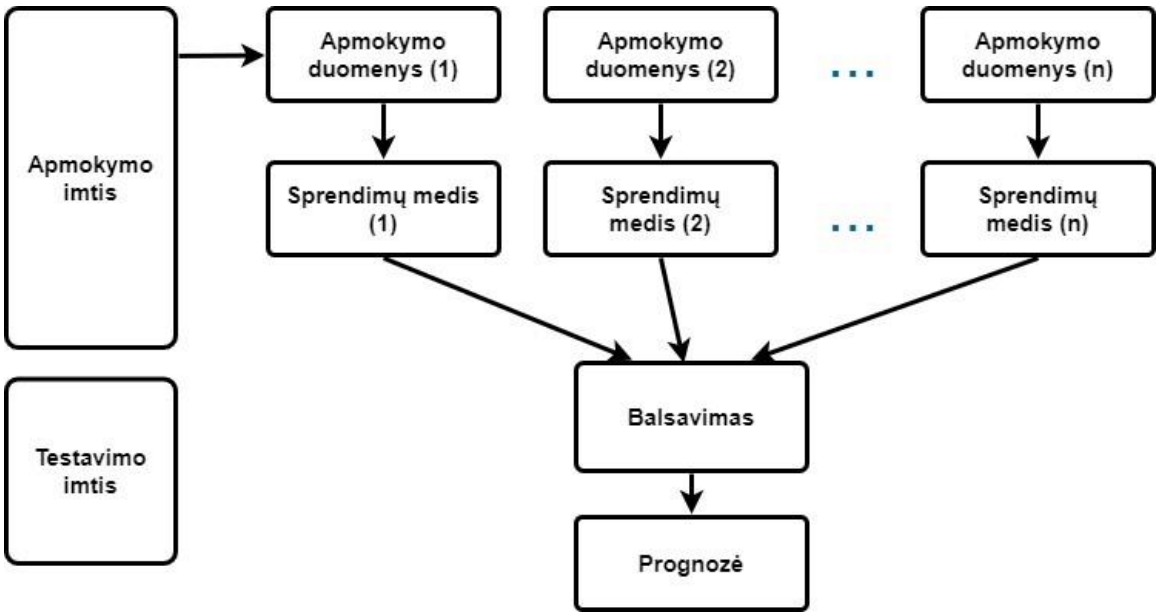
2.3.2. XGBoost

XGBoost yra modifikuota gradiento didinimo versija. Pagrindinis šių algoritmų veikimo principas yra panašus, tačiau yra ir keletas skirtumų:

1. XGBoost modelyje įdiegtos skirtingos reguliavimo technologijos. Tai apima L1 ir L2 reguliavimą, iškritimo (angl. *Dropout*) metodą ir ankstyvą sustabdymą (angl. *Early stopping*). Šis reguliavimo procesas gali sumažinti mokymosi duomenų rinkinio persimokymą arba nepakankamą apsimokymą.
2. Skirtingai nei gradiento didinimas, XGBoost gali lygiagrečiai apdoroti keletą mazgų, tokiu būdu pagreitindamas apmokymo procesą.
3. XGBoost automatiškai supranta priskyrimą (nustato, ar mazgas turi būti dedamas sprendimų medžio dešinėje ar kairėje).
4. XGBoost turi savo integruotą trūkstamų reikšmių tvarkyklę, o gradiento didinimas neturi. Ši tvarkyklė pavadinta „Sparity-aware Split Finding“.
5. XGBoost turi integruotą kryžminį patikrinimą (angl. *Cross validation*), kurį galima naudoti modelio treniravimo metu.

2.3.3. Atsitiktinis miškas

Atsitiktinis miškas (angl. *Random forest*) yra klasifikatorius, kuriame yra daugybė sprendimų medžių (angl. *Decision trees*) įvairiuose duoto duomenų rinkinio pogrupiuose (2.5 pav.). Užuot pasikliovus vienu sprendimų medžiu, atsitiktinis miškas paima kiekvieno medžio prognozę, remdamasis prognozių daugumos balsais, ir tuomet prognozuoja galutinį rezultatą. Didesnis medžių skaičius miške užtikrina didesnį tikslumą ir apsaugo nuo persimokymo.



2.5 pav. Atsitiktinio miško architektūra

2.3.4. Logistinė regresija

Logistinė regresija yra mašininio mokymosi algoritmas, naudojamas prognozuoti tikimybę, kad stebėjimas priklauso vienai iš dviejų galimų klasių.

Logistinės regresijos formulė užrašoma taip:

$$y = \frac{1}{1 + e^{-(w_0 + w_1 x)}}, \tag{8}$$

čia y – prognozuojama tikimybė priklausyti vienai iš dviejų klasių, $\frac{1}{1+e^{-z}}$ – sigmoidinė funkcija, $(w_0 + w_1 x)$ – tiesinis modelis logistinėje regresijoje.

Norint susieti numatomas reikšmes su tikimybėmis, naudojama sigmoidinė funkciją. Ši funkcija bet kurią realią reikšmę susieja su kita reikšme intervale nuo 0 iki 1. Mašininio mokymosi metu naudojama sigmoidinė funkcija, kad prognozes būtų galima susieti su tikimybėmis.

2.3.5. K-artimiausių kaimynų algoritmas

K-artimiausių kaimynų (angl. *K-nearest neighbors, KNN*) algoritmas yra universalus ir plačiai naudojamas mašininiame mokyme. Šio algoritmo pasirinkimą dažniausiai lemia paprastumas ir lengvas panaudojimas. Vienas iš didesnių privalumų yra galimybė apdoroti tiek skaitinius, tiek kategorinius duomenis, todėl tai yra lankstus pasirinkimas įvairių tipų duomenų rinkiniams atliekant klasifikavimo ir regresijos užduotis. KNN algoritmas veikia surandant K artimiausius tam tikro duomenų taško kaimynus pagal atstumo metriką. Tada duomenų taško klasė nustatoma pagal K kaimynų balsų daugumą arba vidurkį.

Dažniausiai naudojamos 3 atstumo metrikos:

– **Euklido atstumas**. Tai ne kas kita kaip dekartinis atstumas tarp dviejų taškų, esančių plokštumoje/hiperplokštumoje. Euklido atstumas taip pat gali būti įsivaizduojamas kaip atkarpos, jungiančios du taškus, į kuriuos reikia atsižvelgti, ilgis:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}\,. \quad (9)$$

– **Manheteno atstumas**. Ši atstumo metrika paprastai naudojama, kai mus domina bendras objekto nukeliautas atstumas, o ne poslinkis:

$$d(x,y) = \sum_{i=1}^{n}|x_i - y_i|\,. \quad (10)$$

– **Minkovskio atstumas**. Galima sakyti, kad Euklido atstumas, taip pat kaip ir Manhatano, yra ypatingi Minkovskio atstumo atvejai. Iš toliau pateiktos formulės galime pasakyti, kad kai p = 2, tada Euklido ir Minkovskio atstumo formulės sutampa, o kai p = 1, tada gaunama Manheteno atstumo formulė:

$$d(x,y) = \left(\sum_{i=1}^{n}(x_i - y_i)^p\right)^{\frac{1}{p}}. \quad (11)$$

Toliau aptariamas KNN algoritmo veikimas:

1. Optimalios *K* vertės pasirinkimas. *K* reiškia artimiausių kaimynų skaičių, į kurį reikia atsižvelgti.
2. Atstumo skaičiavimas. Atstumas apskaičiuojamas tarp kiekvieno duomenų rinkinio duomenų taško ir tikslinio taško.
3. Artimiausių kaimynų paieška. *K* duomenų taškai, kurių atstumai iki tikslinio taško yra mažiausi, identifikuojami kaip artimiausi kaimynai.
4. Klasifikavimo uždavinyje klasių etiketės nustatomos balsų dauguma. Klasė, turinti didžiausią įvertį tarp kaimynų, tampa tiksline duomenų taško klase.

Sukūrus klasifikavimo modelį laukia sekantis žingsnis – kokybės vertinimas. Kitame poskyryje detaliau aptariamos modelių kokybės metrikos, kuriomis naudojantis bus vertinamas modelių tikslumas ir tinkamumas duomenims.

2.4. Klasifikavimo modelių kokybės vertinimas

Bene pats svarbiausias ir atsakingiausias etapas modelių kūrime – sukurto modelio tinkamumo įvertinimas. Pagal gautus rezultatus galima nuspręsti, ar modelis pakankamai gerai klasifikuoja. Be to, tai gali padėti išsirinkti geriausią modelį iš visų galimų arba optimizuoti hiperparametrus atliekant jų derinimą.

2.4.1. Sumaišymo matrica

Sumaišymo matrica (angl. *Confusion matrix*) yra mašininio mokymosi klasifikavimo tikslumo matas. Tai lentelė su numatomų ir faktinių verčių deriniais. Toliau pateiktame 2.6 paveiksle pavaizduota sumaišymo matricos schema.

2.6 pav. Sumaišymo matricos schema

Sumaišymo matricos schemoje naudojamų trumpinių paaiškinimas:

- TP (angl. *True positive*) – teigiami atsiliepimai, kurie priskirti teigiamai klasei;
- FP (angl. *False positive*) – neigiami atsiliepimai, kurie priskirti teigiamai klasei;
- FN (angl. *False negative*) – teigiami atsiliepimai, kurie priskirti neigiamai klasei;
- TN (angl. *True negative*) – neigiami atsiliepimai, kurie priskirti neigiamai klasei.

Pasinaudojant sumaišymo matrica dažniausiai papildomai apskaičiuojamos 4 klasifikavimo metrikos: specifiškumas, jautrumas, bendras tikslumas ir F1 įvertis.

Specifiškumas (angl. *Precision*) paaiškina, kiek teigiamai prognozuotų atvejų iš tikrųjų buvo teisingi. Tikslumas yra naudingas tais atvejais, kai klaidingi teigiami atsiliepimai kelia didesnį susirūpinimą nei klaidingi neigiami. Jis apibrėžiamas kaip teisingų teigiamų atsiliepimų skaičius, padalintas iš bendro prognozuojamų teigiamų atsiliepimų skaičiaus:

$$\text{specifiš}kumas = \frac{TP}{TP + FP}. \tag{12}$$

Jautrumas (angl. *Recall, sensitivity*) paaiškina, kiek faktinių teigiamų atsiliepimų galima teisingai prognozuoti naudojant sukurtą modelį. Jautrumas yra naudingas tais atvejais, kai klaidingi neigiami atsiliepimai kelia didesnį susirūpinimą nei klaidingi teigiami. Ši metrika apibrėžiama kaip tikrų teigiamų atsiliepimų skaičius padalintas iš bendro faktinių teigiamų atsiliepimų skaičiaus:

$$\text{jautrumas} = \frac{TP}{TP + FN}. \tag{13}$$

Bendras tikslumas (angl. *Accuracy*) tai rodiklis, kuris parodo bendrą sukurto modelio tikslumą, neskirstant atsiliepimų į teigiamus ir neigiamus. Jis apibrėžiamas kaip teisingų prognozuotų atsiliepimų skaičius padalintas iš bendro atsiliepimų skaičiaus. Bendras tikslumas apskaičiuojamas:

$$\text{bendras tikslumas} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{14}$$

Kai kurių modelių tikslumas gali siekti net 99 % ir galima manyti, kad modelis veikia labai gerai, tačiau tai ne visada tiesa ir kai kuriose situacijose tai gali klaidinti. Dažniausiai tokia situacija atsiranda tuomet, kai tarp klasių yra didelis disbalansas.

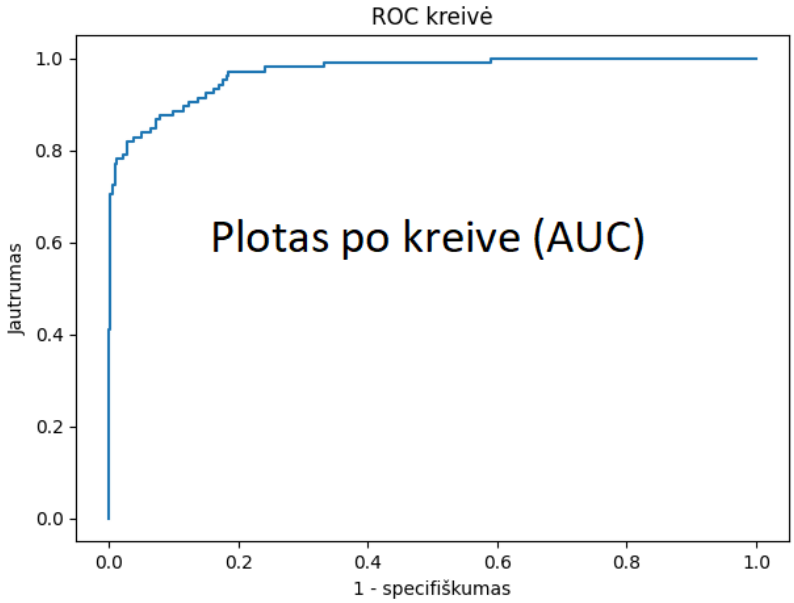
F1 įvertis (angl. *F1 score*) tai bendra metrika, susijusi su specifiškumu ir jautrumu:

$$\text{F1 į}vertis = \frac{2 * specifiškumas * jautrumas}{specifiškumas + jautrumas}. \tag{15}$$

Šios metrikos pagrindinis privalumas tas, kad ją maksimizuojant tuo pačiu metu didėja 2 įverčiai: specifiškumas ir jautrumas.

2.4.2. ROC kreivė ir AUC

ROC (angl. *Receiver operating characteristic*) kreivė (2.7 pav.) parodo specifiškumo ir jautrumo sąryšį. Jei modelis klasifikuoja atsitiktinai, tuomet ROC kreivė yra tiesės formos. Kuo kreivė labiau išlinkusi į viršų, tuo modelio klasifikavimas yra tikslesnis. Jei kreivė išlinksta į apačią, tai reiškia, kad modelis atlieka daugiau klaidingų klasifikavimų nei teisingų. Didesnė *Ox* ašies reikšmė rodo geresnį klaidingų teigiamų (FP) atsiliepimų klasifikavimą nei tikrų neigiamų (TN), o didesnė *Oy* ašies reikšmė rodo didesnį tikrų teigiamų (TP) atsiliepimų skaičių nei klaidingų neigiamų (FN).



2.7 pav. ROC kreivės ir AUC grafikas

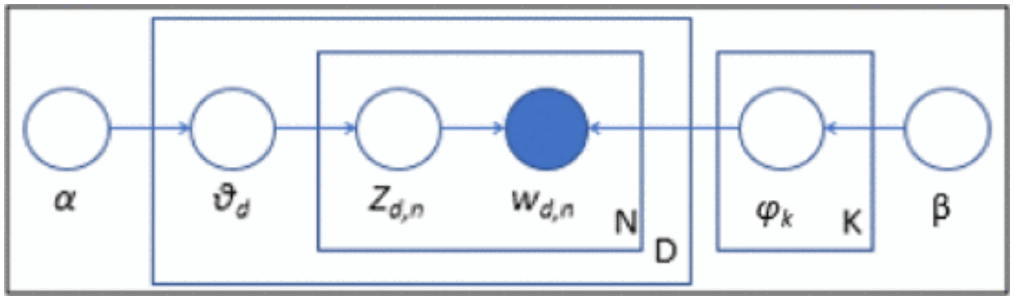
Analizuojant anksčiau pateiktą grafiką, galima pastebėti ploto po kreive (angl. *AUC - Area under curve*) rodiklį, kuris parodo modelio našumą skirtinguose slenksčiuose tarp teigiamų ir neigiamų klasių. Tai reiškia, jei AUC yra lygus 1, klasifikatorius gali puikiai atskirti visus teigiamus ir neigiamus atsiliepimus.

Klasifikavimas padeda tais atvejais, kai atsiliepimų klasės yra iš anksto žinomos (šiuo atveju tai yra teigiami ir neigimai atsiliepimai). Tačiau kai kuriais atvejais klasės nėra žinomos arba jose nėra reikalingos informacijos. Tokiais atvejais galima pasinaudoti temos modeliavimu, kuris detaliau aptariamas kitame poskyryje.

2.5. Temos modeliavimas

Panaudojus neprižiūrimą mašininį mokymąsi galima išgauti atsiliepimo temą. Vienas iš tokių metodų yra Atviras Dirichlet pasiskirstymas (angl. *Latent dirichlet allocation, LDA*). Pagrindinis LDA uždavinys – surasti temą, kuri teksto rinkinyje pasiskirsto su didžiausia tikimybe. Verta paminėti, kad LDA naudoja žodžių krepšelį (angl. *Bag of words*), kuris neatsižvelgia į žodžių tvarką. Toliau pateiktame 2.8 paveiksle pateiktas vizualus LDA metodo modelis. Modelyje esančios reikšmės: D – atsiliepimų skaičius, N – žodžių skaičius atsiliepime. Taip pat yra parametrai, kuriuos taip pat būtina nurodyti: K – temų skaičius, α – temos pasiskirstymas atsiliepime ir β – žodžio pasiskirstymas

temoje. Atsiliepime žodžiui priskirta tema apibrėžiama kaip $Z_{d,n}$, o $W_{d,n}$ – dokumente pastebėti žodžiai.



2.8 pav. Grafinis LDA metodo modelis

LDA turi du skirtingus procesus: generatyvinį (angl. Generative) ir išvadų (angl. Inference). Generatyvinis procesas taikomas tuomet, kai tokie parametrai kaip žodžių pasiskirstymas temoje (φ_k) ir temos proporcija kiekvienam atsiliepimui ($\theta_{d,n}$), yra žinomi iš anksto. Tuo tarpu išvadų procesas taikomas norint nustatyti tuos parametrus ir žodžių pasiskirstymą ($\omega_{d,n}$) tarp temų, remiantis turimu atsiliepimu. Taikant išvadų procesą, svarbiausia užduotis yra apsiskaičiuoti φ_k ir $\theta_{d,n}$ reikšmes pagal tokias formules:

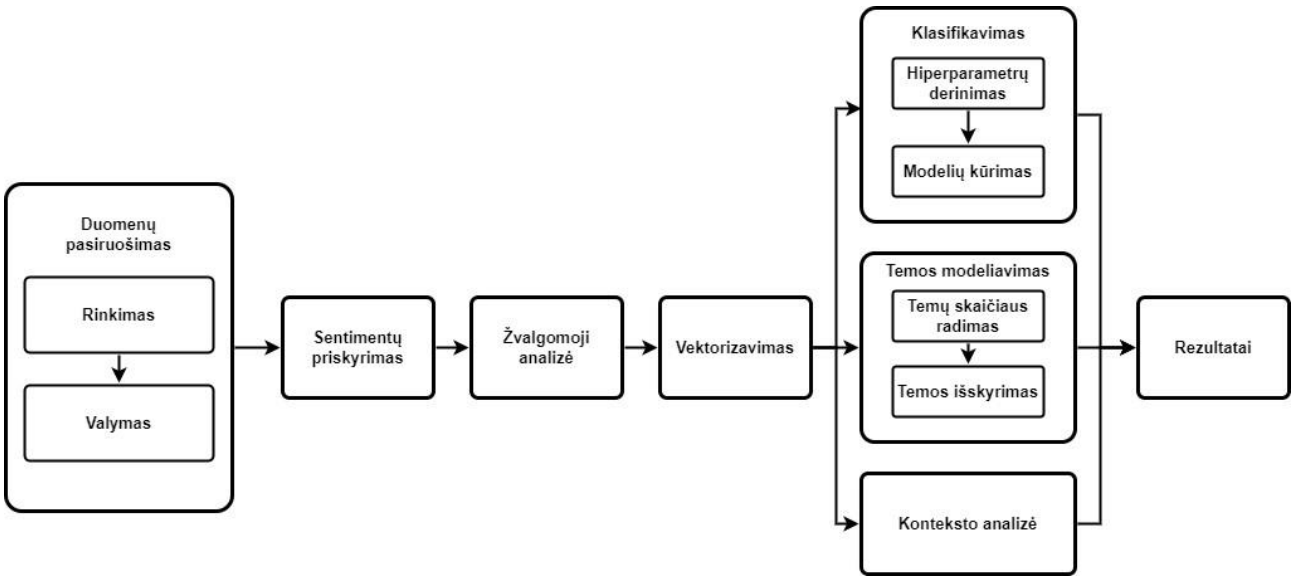
$$\varphi_k = p(\omega = t|z = k) = \frac{n_{t,k} + \beta_t}{\sum_{t=1}^{V} n_{t,k} + \beta_t}, \tag{16}$$

$$\theta_{d,n} = p(z = k|d) = \frac{n_{d,k} + \alpha_k}{\sum_{k=1}^{K} n_{d,k} + \alpha_k}, \tag{17}$$

čia $n_{t,k}$ – temai priskirtų žodžių skaičius, $n_{d,k}$ – temai priskirtų žodžių skaičius dokumente, o V – skirtingų žodžių skaičius dokumente.

Siekiant parinkti teisingus parametrus LDA metodui, pasinaudota koherentiniu (angl. *Cogerence*) C_v įverčiu. Šis įvertis yra pagrįstas slankiojančiu langu (angl. *Sliding window*), vieno pagrindinio žodžio segmentavimu ir netiesiogine patvirtinimo priemone, kuri naudoja normalizuotą taškinę abipusę informaciją (NPMI) ir kosinuso panašumą.

Prieš pereinant į tyrimo rezultatų skyrių, sudarytas vizualus metodologijos apibendrinimas (2.9 pav.), kuriame glaustai pateikiamas visas tyrimo procesas.



2.9 pav. Tyrime naudojamos metodologijos apibendrinimas

Kitame skyriuje detaliau pakomentuojamas kiekvienas etapas bei rezultatai, kuriuos pavyko gauti.

3. Tyrimo rezultatai

Šiame skyriuje aprašomi atsiliepimų analizės etapai bei rasti klientų praradimą lemiantys veiksniai.

3.1. Duomenų pasiruošimas

Klientų praradimą lemiančių veiksnių analizei buvo pasirinktos 5 interneto svetainės, kuriose surinkti atsiliepimai apie 3 Lietuvos telekomunikacijų įmones (3.1 lentelė). Šiai užduočiai pasitelkta Python programavimo kalba bei keletas bibliotekų. Sudėtingiausias duomenų rinkimas buvo iš „Facebook“ socialinio tinklo, naudojant *facebook-scraper* biblioteką. „Facebook“ socialinis tinklas taiko gana griežtą politika kalbant apie duomenų rinkimą, dėl šios priežasties gauti didelius duomenų rinkinius yra labai sudėtinga, o kartais net neįmanoma. Rastos „Facebook“ grupės, susijusios su telekomunikacijų įmonėmis, yra gana naujos ir kol kas neturi daug įrašų, tačiau jų populiarumas gana greitai auga, ir tikėtina, kad ateityje jose bus galima rasti daug naudingos informacijos. Tuo tarpu iš kitų interneto svetainių atsiliepimus pavyko išgauti gana lengvai ir greitai pasinaudojant *BeautifulSoup* biblioteka. Šios bibliotekos pagalba buvo nuskaitomi pasirinktų svetainių HTML kodai ir tuomet, atsirinkus reikiamas žymes, paimama norima informacija.

3.1 lentelė. Analizuojamų duomenų šaltiniai

Šaltinis	**Telekomunikacijų įmonė**	**Atsiliepimų kiekis**	**Seniausio atsiliepimo data**
Facebook (grupė: „Tele2 Lietuva skundai ir atsiliepimai“)	Tele2	84	2022-10-05
Facebook (grupė: „Telia Lietuva skundai ir atsiliepimai“)	Telia	26	2023-01-14
Facebook (grupė: „Bitė. Lietuva. Nusiskundimai“)	Bitė	29	2022-03-07
Rekvizitai.lt	Tele2	5 594	2012-05-01
	Telia	1 175	2017-12-19
	Bitė	4 438	2012-06-22
Atsiliepimai.lt	Tele2	64	2017-04-21
	Telia	40	2017-04-13
	Bitė	70	2017-04-21
Imones.lt	Tele2	30	2017-07-09
	Telia	139	2014-09-26
	Bitė	129	2014-06-05
Matuokle.lt	Tele2	302	2010-01-01
	Telia	508	2010-01-11
	Bitė	43	2010-01-11

Surinkus klientų atsiliepimus, būtina juos apdoroti bei pašalinti nereikalingą informaciją, kuri gali turėti neigiamos įtakos tyrimo rezultatams. Pašalinus nereikalingą informaciją, atliktas lemavimas, kuris padėjo gerokai sumažinti skirtingų žodžių skaičių, suvienodinant jų gramatines formas. Šiaip užduočiai atlikti panaudota *simplemma* biblioteka, kuri gana tiksliai sugeba apdoroti net lietuvių kalbą. Kai kurių žodžių lemavimas ir gramatinių klaidų taisymas buvo atliktas rankiniu būdu, siekiant

kuo tikslesnių rezultatų. Paskutinis žingsnis yra teksto skaidymas dalimis (angl. *Tokenization*), kuris suskaido tekstą į mažesnes dalis, šiuo atveju į žodžius. Toliau esančioje 3.2 lentelėje pateikti atsiliepimai prieš teksto apdorojimą ir po jo.

3.2 lentelė. Atsiliepimų apdorojimo pavyzdys

Originalus atsiliepimas	**Atsiliepimas po apdorojimo**
Prastas interneto ryšys Trakų raj. Karalūnų k.	[prastas, internetas, ryšys]
Nuoširdziai niekam nerekomenduoju rinktis šio tinklo.	[niekam, nerekomenduoti, rinktis, tinklas]
Melagiai ir apgavykai. Visais būdai venkite jų namų interneto	[melagis, būdas, vengti, namas, internetas]
Interneto paslaugos tragiškos.	[internetas, paslauga, tragiškas]

Sutvarkius atsiliepimus bei pašalinus nenaudingą informaciją, galima pereiti prie kito žingsnio – sentimentų priskyrimo.

3.2. Sentimentų priskyrimas

Sentimentų priskyrimas atliktas analizuojant kiekvieną atsiliepimą atskirai ir rankiniu būdu identifikuojant sentimentą. Pasirinkti 3 sentimentų tipai: teigiamas, neigiamas ir neutralus. Sentimentų pasiskirstymas pateiktas toliau esančioje 3.3 lentelėje.

3.3 lentelė. Sentimentų pasiskirstymas

Sentimentas	**Aprašymas**	**Atsiliepimų skaičius**
Teigiamas	Vartotojo atsiliepimas, kuriame išreiškiamas pasitenkinimas naudojamomis paslaugomis.	519
Neigiamas	Vartotojo atsiliepimas, kuriame išreiškiamas nepasitenkinimas naudojamomis paslaugomis.	7 561
Neutralus	Aiškios emocijos neturintis atsiliepimas.	913
Iš viso		**8 825**

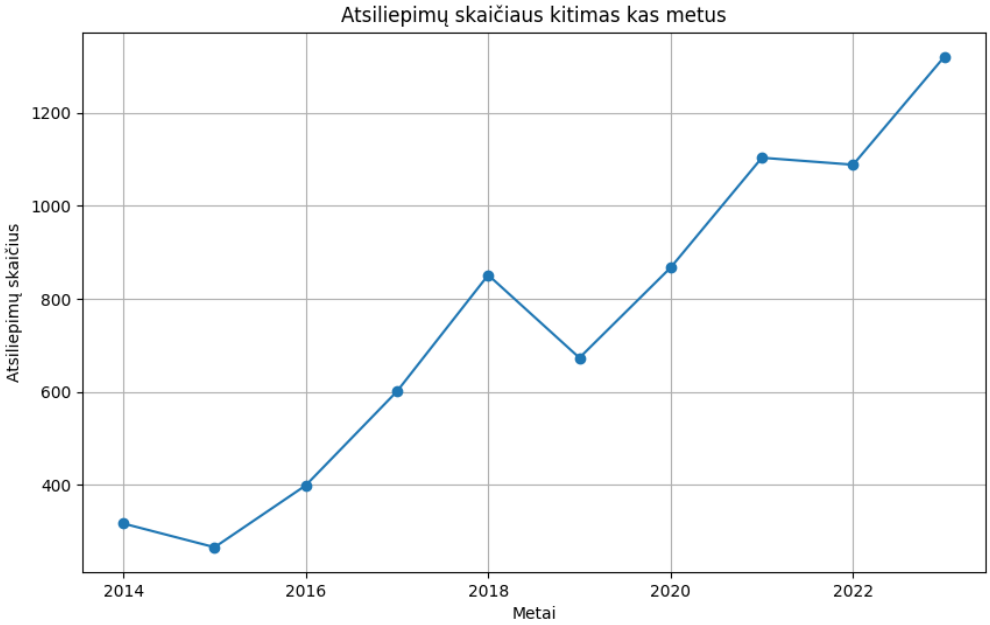
Atsiliepimai, kuriuose teigiamai pasisakoma apie įmonės teikiamas paslaugas, įvardinti kaip turintys teigiamą sentimentą. Jei identifikuojama, kad kliento atsiliepimas yra neigiamas, tai reiškia, kad klientas yra nepatenkintas įmonės teikiamomis paslaugomis ir yra gana didelė tikimybė, kad šis klientas artimiausiu laiku atsisakys paslaugų ir pasirinks kitą telekomunikacijas teikiančia įmonę. Visi kiti atsiliepimai, kuriuose nerasta aiškios emocijos, įvardinti kaip neutralūs.

Tolesnėje analizėje nuspręsta tirti tik neigiamus ir teigiamus atsiliepimus. Neutraliuose atsiliepimuose dažnu atveju neatsiskleidžia kliento pozicija arba pateikiamas visiškai su telekomunikacijomis nesusijusi informacija. Dėl šių priežasčių šie atsiliepimai buvo įvardinti kaip neaktualūs, analizuojant vartotojų praradimą lemiančius veiksnius.

3.3. Žvalgomoji analizė

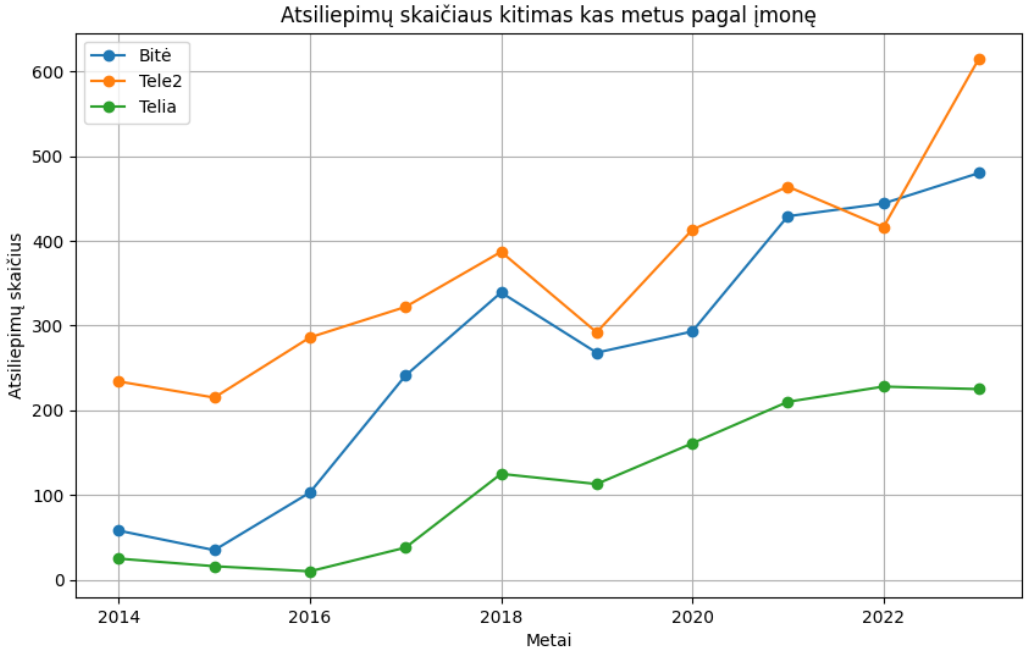
Turint atsiliepimus bei jiems priskirtus sentimentus, verta detaliau paanalizuoti turimus duomenis įvairiais pjūviais. Kadangi pirmieji atsiliepimai apie telekomunikacijų įmones rasti dar 2010 metais, kuomet vyravo visai kitos naudojimosi mobiliuoju ryšiu tendencijos nei dabar, todėl nuspręsta analizuoti tik paskutinio dešimtmečio atsiliepimus, tai yra tuos, kurie parašyti ne anksčiau nei 2014 metais. Tikslingiausia būtų analizuoti tik paskutinių kelerių metų atsiliepimus, tačiau dėl nedidelio

duomenų rinkinio nuspręsta tyrime naudoti ilgesnį laikotarpį. Pirmiausia įvertinama, kaip kito atsiliepimų skaičius (3.1 pav.). Galima pastebėti, kad vartotojai internete palieka vis daugiau atsiliepimų ir kiekvienais metais jų skaičius auga. Pirmaisiais metais atsiliepimų buvo parašoma iki 400, o paskutiniais keliais metais jų skaičius viršija 1000. Per praėjusius 2023 metus internete pavyko rasti daugiau nei 1300 atsiliepimų, susijusių su trimis didžiausiomis telekomunikacijų įmonėmis Lietuvoje.



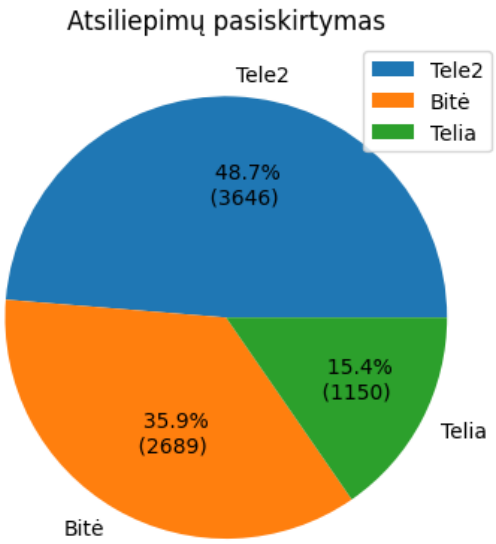
3.1 pav. Atsiliepimų skaičiaus kitimo grafikas

Lyginant atsiliepimų skaičiaus kitimą kiekvienai įmonei atskirai (3.2 pav.), galima pastebėti, kad visoms analizuotoms įmonėms klientai kiekvienais metais palieka vis daugiau atsiliepimų. Ypač tai pastebima Bitės ir Tele2 grafikuose, tuo tarpu Telia paskutiniais metais sulaukia gerokai mažiau atsiliepimų negu kiti operatoriai ir augimas taip pat yra menkesnis.



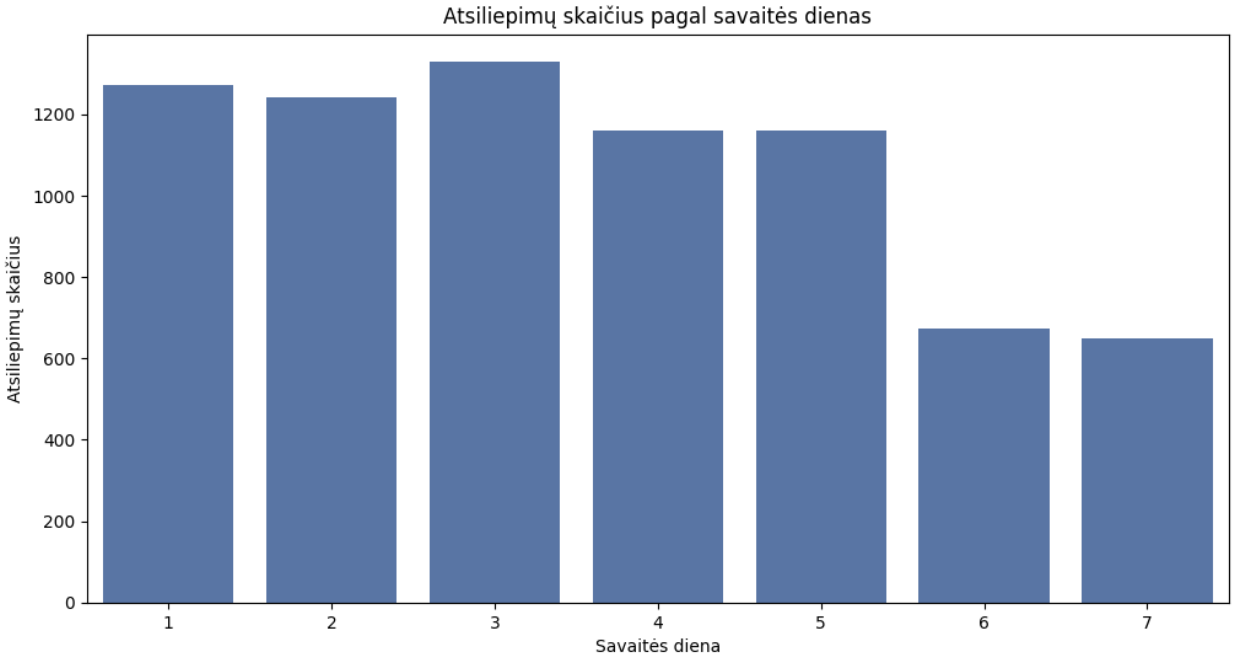
3.2 pav. Atsiliepimų skaičiaus kitimo pagal įmonę grafikas

Atsiliepimų pasiskirstymo pagal įmones skritulinėje diagramoje (3.3 pav.) taip pat pastebima, kad mažiausiai atsiliepimų sulaukia Telia. Tuo tarpu beveik pusė visų surinktų atsiliepimų yra apie Tele2 įmonę ir jos teikiamas paslaugas. Tikėtina, kad atsiliepimų skaičius tiesiogiai koreliuoja su bendru klientų skaičiumi. Deja, bet viešai prieinamos informacijos apie kiekvienos telekomunikacijų įmonės klientų skaičių nepavyko rasti.



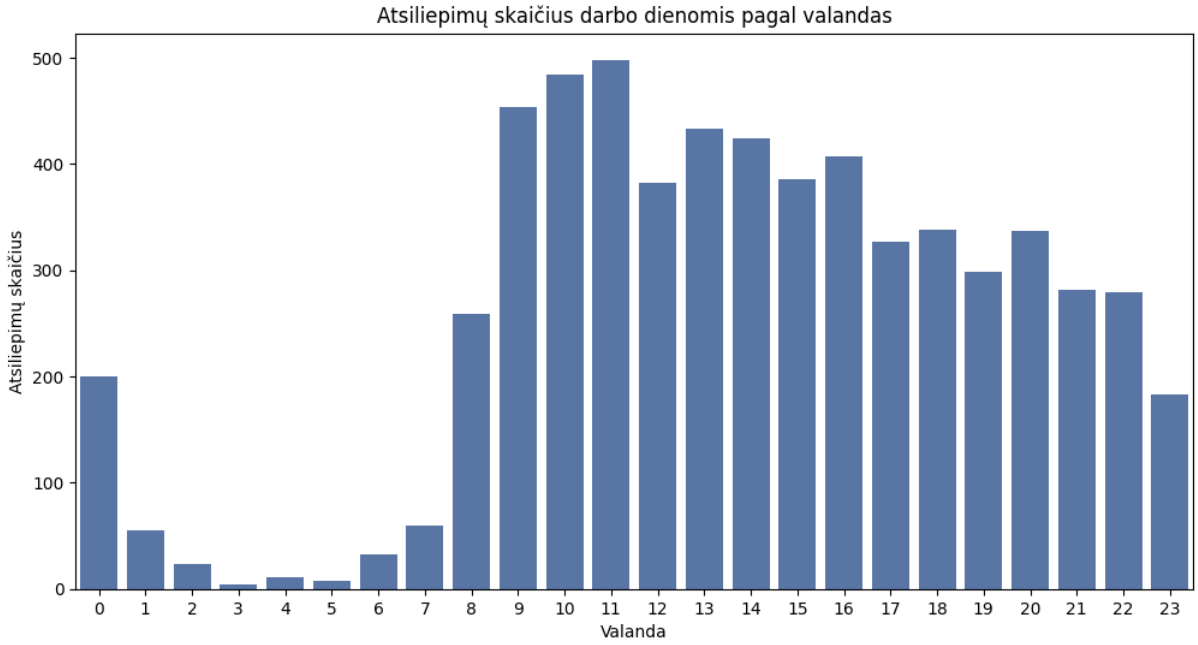
3.3 pav. Atsiliepimų pasiskirstymo pagal įmones skritulinė diagrama

Analizuojant atsiliepimų skaičių pagal savaitės dienas (3.4 pav.), pastebėta gana įdomi tendencija. Darbo dienomis parašoma maždaug 2 kartus daugiau atsiliepimų negu savaitgaliais. Galima daryti prielaidą, kad savaitgaliais daugiau laiko skiriama kitoms veikloms ir klientai yra mažiau suinteresuoti išreikšti savo nuomonę internete.



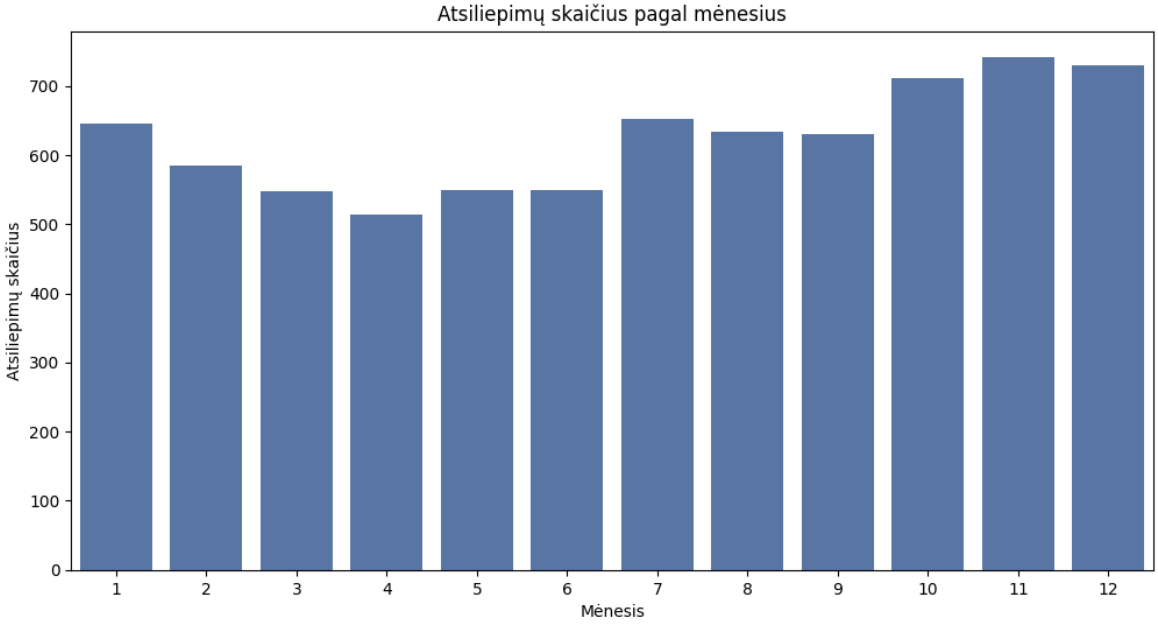
3.4 pav. Atsiliepimų skaičiaus pagal savaitės dienas grafikas

Kaip pastebėta ankstesniame grafike, darbo dienomis klientai aktyviai reiškia nuomonę apie telekomunikacijų įmones, tačiau įdomu išsiaiškinti, kokiomis darbo dienų valandomis atsiliepimų parašoma daugiausiai. Kaip matoma toliau pateiktame grafike (3.5 pav.), klientai aktyviausiai savo nuomonę reiškia dienos viduryje, maždaug 11–16 valandomis. Įprastai šiomis valandomis yra dirbama, todėl galima daryti prielaidą, kad didžioji dalis atsiliepimų yra parašomi būtent darbe.



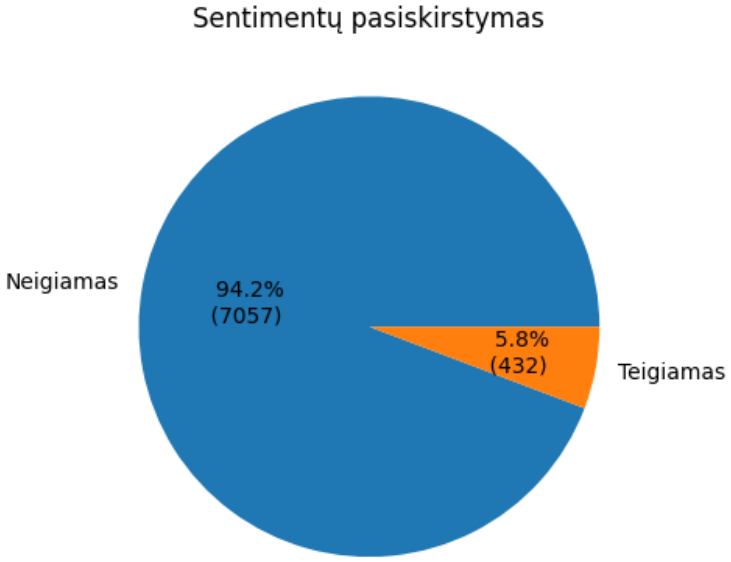
3.5 pav. Atsiliepimų skaičiaus darbo dienomis pagal valandas grafikas

Be to, nuspręsta patikrinti, kuriais mėnesiais parašoma daugiausiai atsiliepimų. Kaip galima pastebėti iš grafiko (3.6 pav.), daugiausiai atsiliepimų parašoma spalio, lapkričio ir gruodžio mėnesiais. Tai galima sieti su rudenišku oru ir ilgais vakarais, kuomet vartotojai yra mažiau užimti ir gali daugiau laiko skirti atsiliepimų rašymui. Taip pat tam įtakos gali turėti išaugęs naudojimasis telekomunikacijų paslaugomis bei džiaugsmai ar problemos, su kuriomis susiduriama.



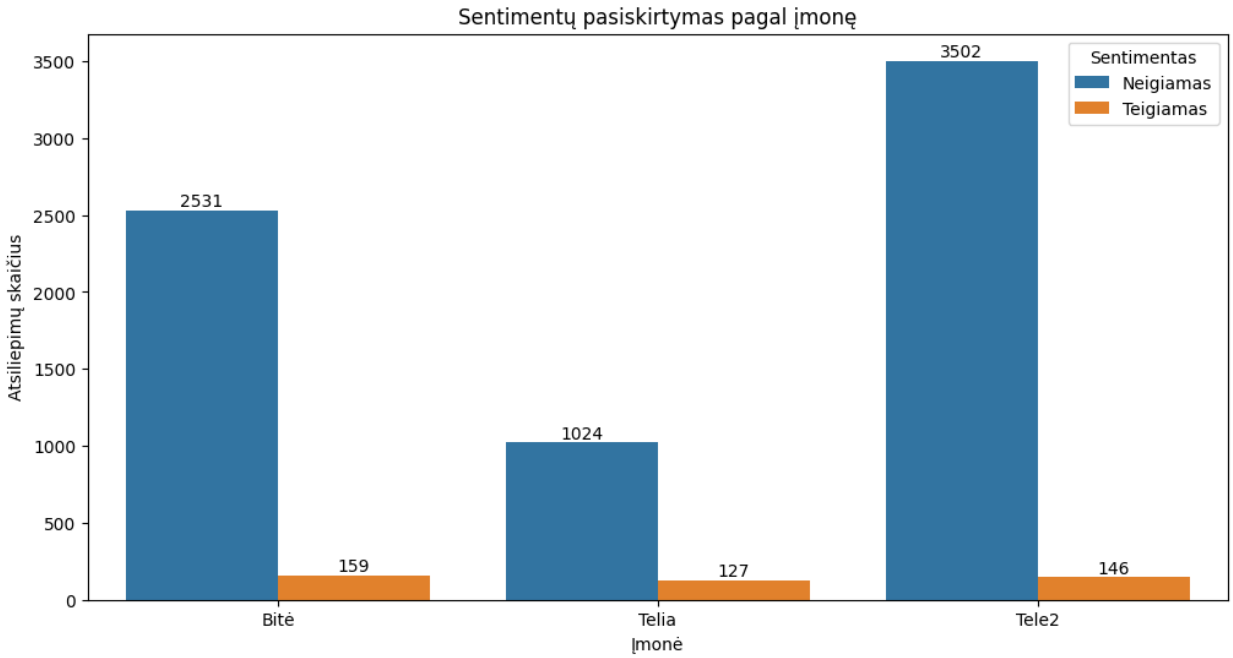
3.6 pav. Atsiliepimų skaičiaus pagal mėnesius grafikas

Renkant atsiliepimus iš įvairių interneto svetainių, pastebėta, kad klientai palieka gerokai daugiau neigiamų atsiliepimų negu teigiamų. Kad būtų galima lengviau palyginti teigiamų bei neigiamų atsiliepimų santykį, proporcija pateikta skritulinėje diagramoje (3.7 pav.). Teigiamų atsiliepimų klientai nėra linkę rašyti, todėl teigiami atsiliepimai sudaro tik maždaug 6 % visų atsiliepimų.



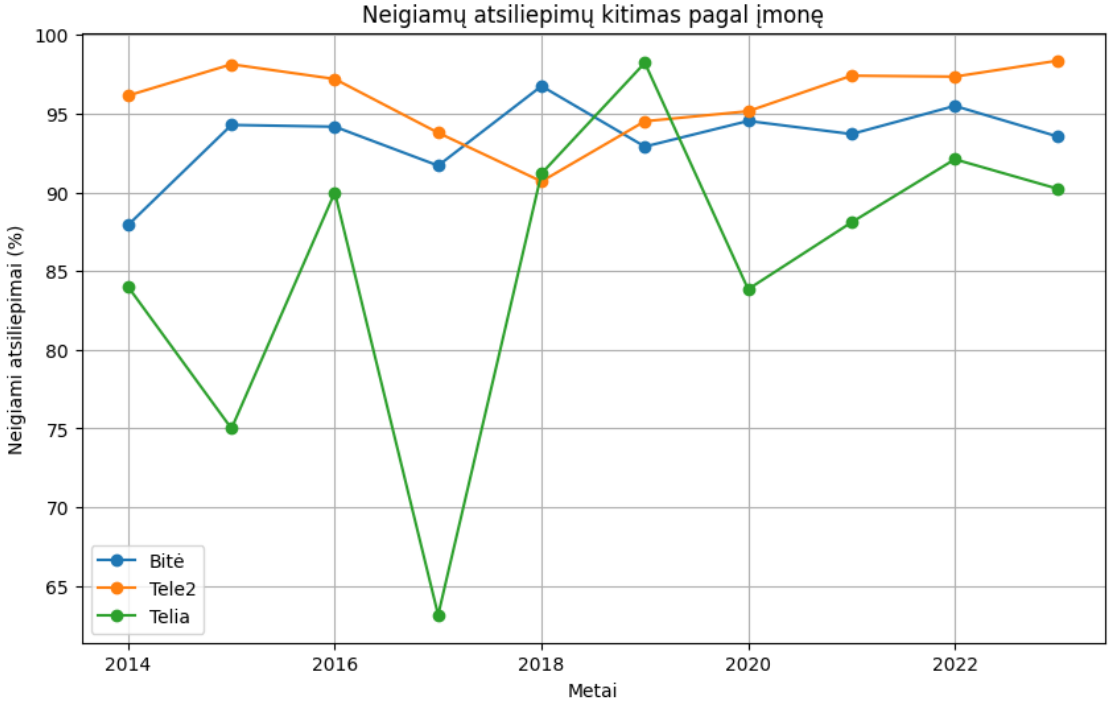
3.7 pav. Sentimentų pasiskirstymo skritulinė diagrama

Lyginant teigiamus ir neigiamus atsiliepimus kiekvienai telekomunikacijų įmonei (3.8 pav.), pastebėta, kad visos 3 analizuotos įmonės turi labai panašų skaičių teigiamų atsiliepimų. Tačiau pastebimas labai ženklus neigiamų atsiliepimų skirtumas. Iš visų pateiktų įmonių mažiausiai neigiamų atsiliepimų turi Telia, tuo tarpu Bitė ją lenkia maždaug dvigubai, o Tele2 net 3 kartus.



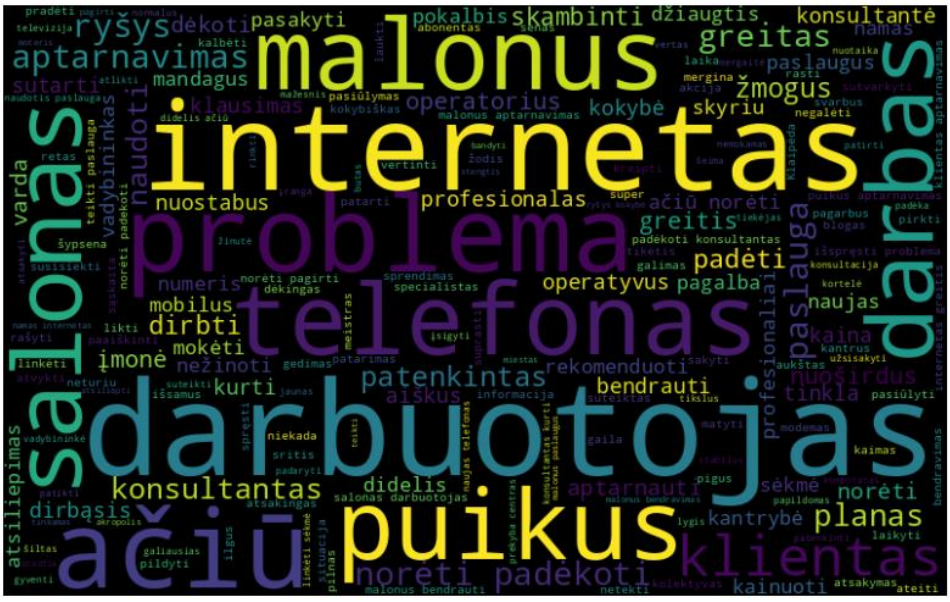
3.8 pav. Sentimentų pasiskirstymo stulpelinė diagrama

Analizuojant teigiamų ir neigiamų atsiliepimų santykį laike (3.9 pav.), Bitė ir Tele2 atveju nepastebėta jokių ženklių pokyčių, maždaug 90–95 % visų atsiliepimų sudaro būtent neigiami. Tuo tarpu Telia teigiamų ir neigiamų atsiliepimų santykis gana ženkliai svyruoja, tam įtakos gali turėti gerokai mažesnis bendras atsiliepimų skaičius.



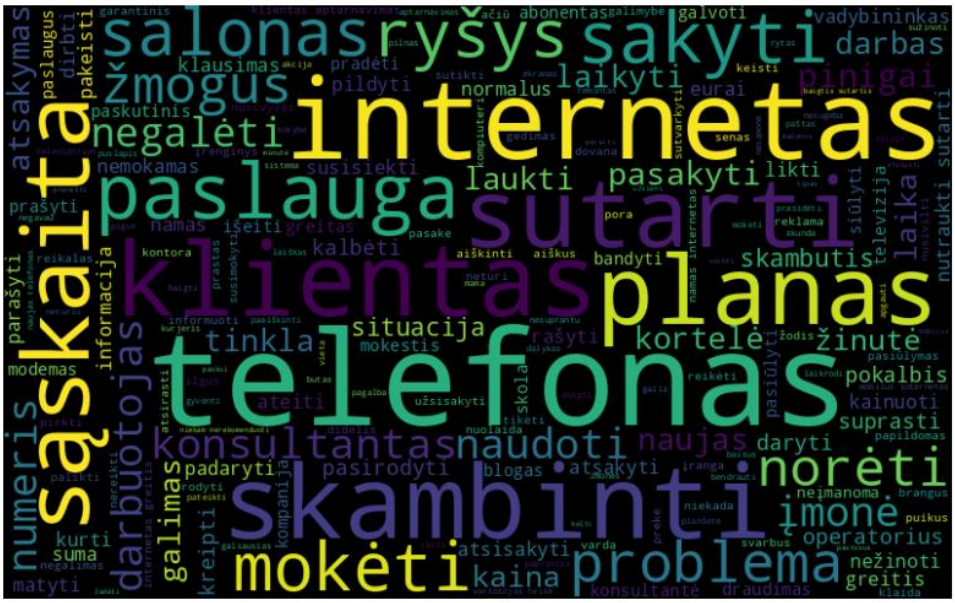
3.9 pav. Neigiamų atsiliepimų kitimo pagal įmonę grafikas

Siekiant išsiaiškinti, kokie žodžiai dominuoja teigiamuose atsiliepimuose, nuspręsta sudaryti žodžių debesį (3.10 pav.). Kaip ir buvo galima tikėtis, pastebimi teigiamas emocijas reiškiantys žodžiai, tokie kaip „malonus“, „puikus“ ir „ačiū“. Tačiau daug yra ir bendresnių žodžių, tokių kaip „internetas“, „darbuotojas“ ar „salonas“.



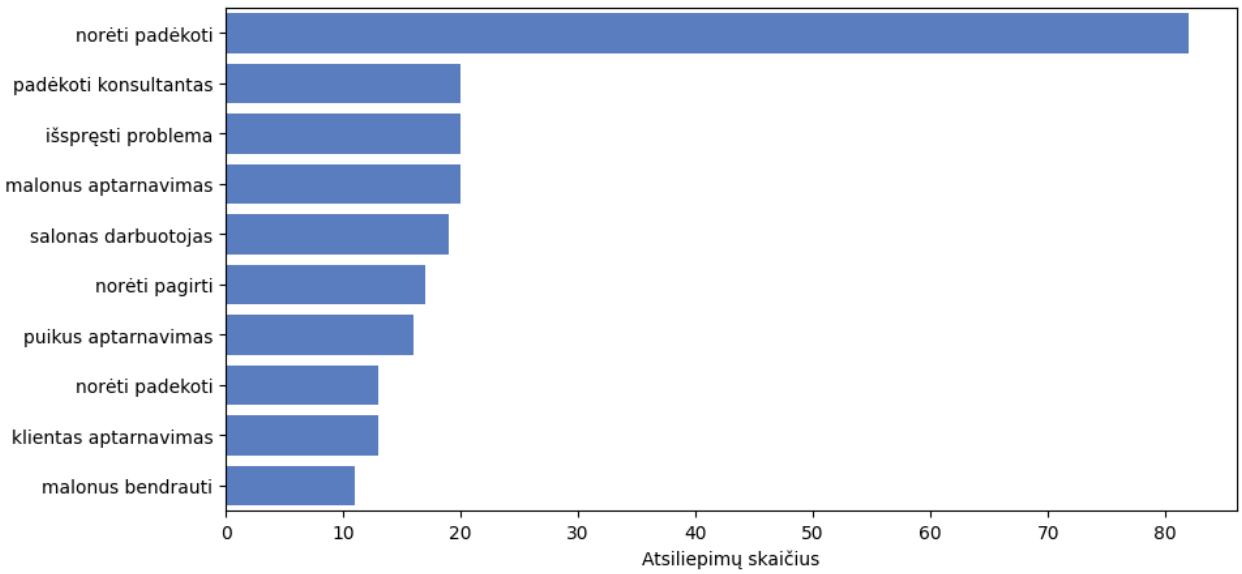
3.10 pav. Teigiamų atsiliepimų žodžių debesis

Tuo tarpu neigiamų atsiliepimų žodžių debesyje (3.11 pav.) gerokai sunkiau suprasti atsiliepimo emociją, nes dominuoja bendriniai žodžiai, tinkantys tiek teigiamam, tiek neigiamam sentimentui.



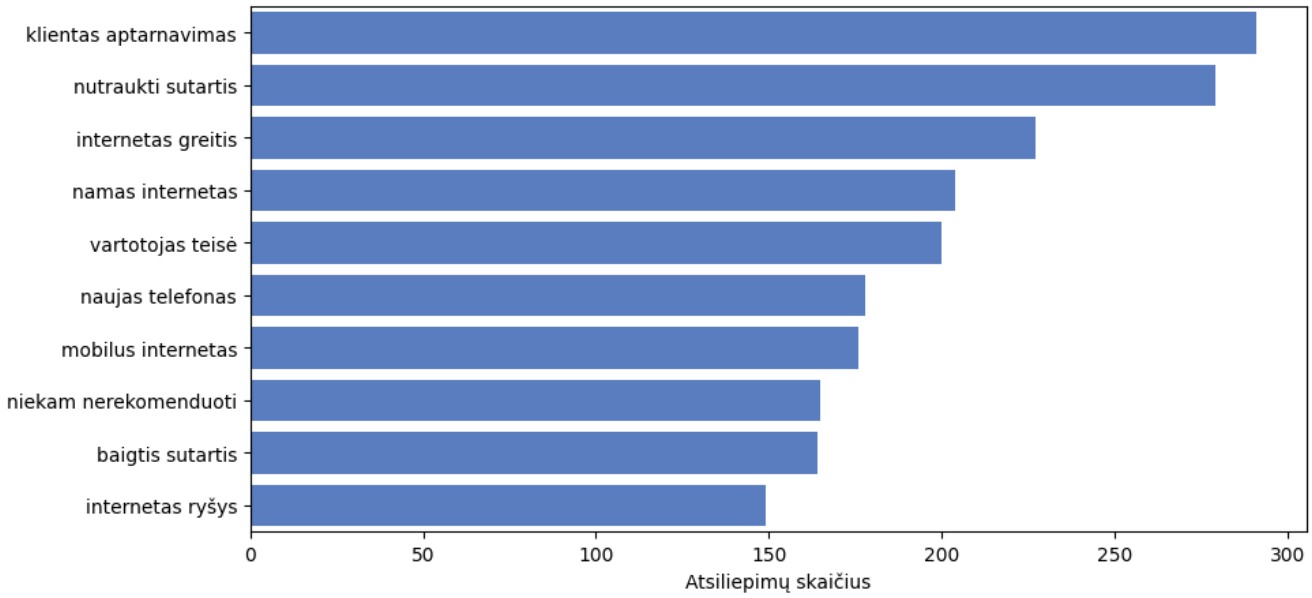
3.11 pav. Neigiamų atsiliepimų žodžių debesis

Kadangi analizuojant pavienius žodžius nepavyko įžvelgti aiškios atsiliepimų emocijos, todėl nuspręsta įtraukti ne pavienius žodžius, o jų junginius. Toliau pateiktame grafike (3.12 pav.) galima nesunkiai suprasti, kad čia pasirinkti būtent teigiami atsiliepimai, nes dominuoja tokios frazės kaip „norėti padėti“, „padėkoti konsultantas“ ir „išspręsti problema“.



3.12 pav. Teigiamuose atsiliepimuose dominuojantys dviejų žodžių junginiai

Žvelgiant į neigiamuose atsiliepimuose dominuojančius žodžių junginius (3.13 pav.), galima įžvelgti kelias frazes, kurios būdingos tik neigiamiems atsiliepimams: „nutraukti sutartis“ ir „niekam nerekomenduoti.“ Visos kitos frazės gana aiškiai identifikuoja veiksnius, kuriais klientai yra neaptenkinti. Frazę „klientas aptarnavimas“ galima sieti su prastu aptarnavimu, o „internetas greitis“ su prastu interneto greičiu.



3.13 pav. Neigiamuose atsiliepimuose dominuojantys dviejų žodžių junginiai

Žvalgomoji duomenų analizė atlikta, galima pareiti prie kito žingsnio, kuris susijęs su atsiliepimų klasifikavimu pagal sentimentus.

3.4. Klasifikavimo modeliai

Pirmasis žingsnis norint klasifikuoti atsiliepimus – duomenų vektorizavimas. Vektorizavimui atlikti pasirinkti 4 metodai: žodžių krepšelis, termino dažnis – atvirkštinis dokumento dažnis, word2vec (skip-gram) ir word2vec (CBOW). Kuriant word2vec modelius bandytas skirtingas žodžių vektoriaus dydis, tačiau įvertinus gautus rezultatus nuspręsta naudoti standartinę reikšmę, kuri lygi 100. Kai tekstas vektorizuotas, tuomet galima pradėti modelių kūrimą. Modelių kūrimui panaudotos sklearn ir xgboost bibliotekos. Didžioji dalis naudojamų modelių turi parametrų derinimo galimybę, todėl nuspręsta pasinaudoti šia galimybe ir išgauti kuo geresnius klasifikavimo rezultatus. Kryžminiame patikrinime naudotos 5 dalys, o visi derinti modelių parametrai pateikiami 3.4 lentelėje.

3.4 lentelė. Derinti modelių parametrai bei geriausios jų reikšmės

Modelis	**Parametras**	**Geriausia reikšmė**	**Kryžminio patikrinimo dalys**
XGBoost (BOW)	max_depth	3	5
	min_child_weight	2	
	n_estimators	300	
XGBoost (TF-IDF)	max_depth	3	
	min_child_weight	2	
	n_estimators	200	
XGBoost (SG)	max_depth	3	
	min_child_weight	4	
	n_estimators	400	
XGBoost (CBOW)	max_depth	3	

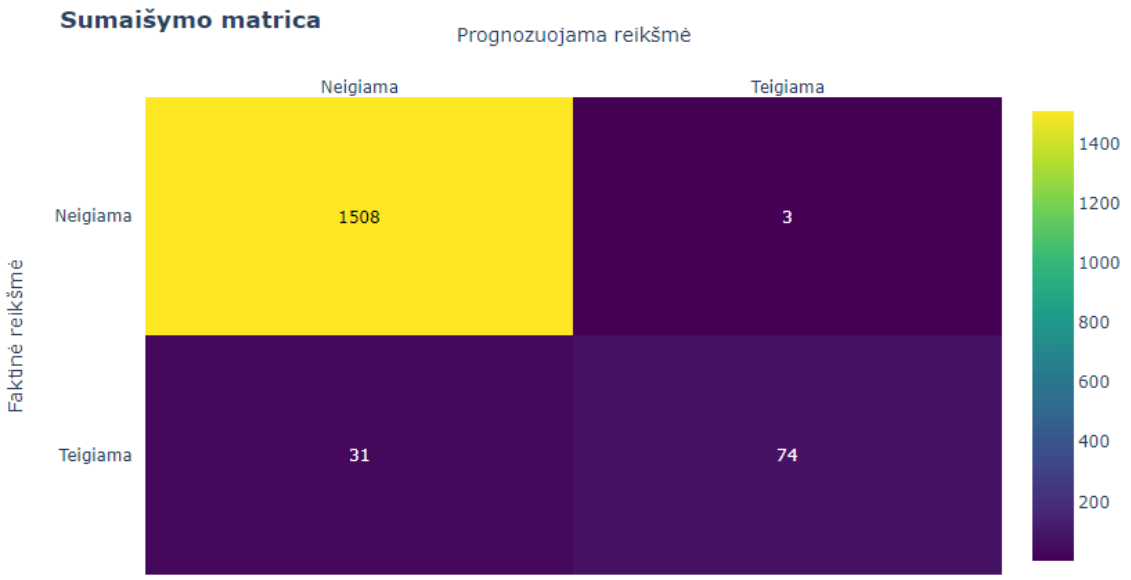
	min_child_weight	2	
	n_estimators	500	
K-artimiausių kaimynų (BOW)	n_neighbors	3	5
K-artimiausių kaimynų (TF-IDF)		11	
K-artimiausių kaimynų (SG)		3	
K-artimiausių kaimynų (CBOW)		5	
Atsitiktinis miškas (BOW)	n_estimators	200	5
	min_samples_leaf	1	
	max_depth	40	
Atsitiktinis miškas (TF-IDF)	n_estimators	300	
	min_samples_leaf	1	
	max_depth	40	
Atsitiktinis miškas (SG)	n_estimators	300	
	min_samples_leaf	1	
	max_depth	20	
Atsitiktinis miškas (CBOW)	n_estimators	300	
	min_samples_leaf	3	
	max_depth	20	
Gradiento didinimas (BOW)	learning_rate	0,1	5
	max_depth	7	
Gradiento didinimas (TF-IDF)	learning_rate	0,1	
	max_depth	5	
Gradiento didinimas (SG)	learning_rate	0,1	
	max_depth	3	
Gradiento didinimas (CBOW)	learning_rate	0,1	
	max_depth	3	

Panaudojus 4 skirtingus vektorizavimo ir 5 klasifikavimo metodus, iš viso sukurta 20 modelių bei įvertintas jų klasifikavimo tikslumas. Toliau pateiktoje 3.5 lentelėje pavaizduoti visų sukurtų modelių klasifikavimo rezultatai. Modelių kūrimui panaudota 80 % visos duomenų imties, o testavimui – 20 %. Kadangi tarp klasifikuojamų klasių yra didelis disbalansas, todėl vertinant modelių kokybę didžiausias dėmesys skirtas F1 įverčiui. Kai kuriems modeliams didelis disbalansas buvo rimtas iššūkis ir F1 įvertis siekia tik 76 %, tačiau yra ir tokių, kurie gana gerai susidorojo su užduotimi ir pasiekė net 90 %. Geriausias klasifikavimo rezultatas pasiektas naudojant TF-IDF vektorizavimą ir logistinės regresijos klasifikavimo metodą.

3.5 lentelė. Klasifikavimo modelių rezultatai

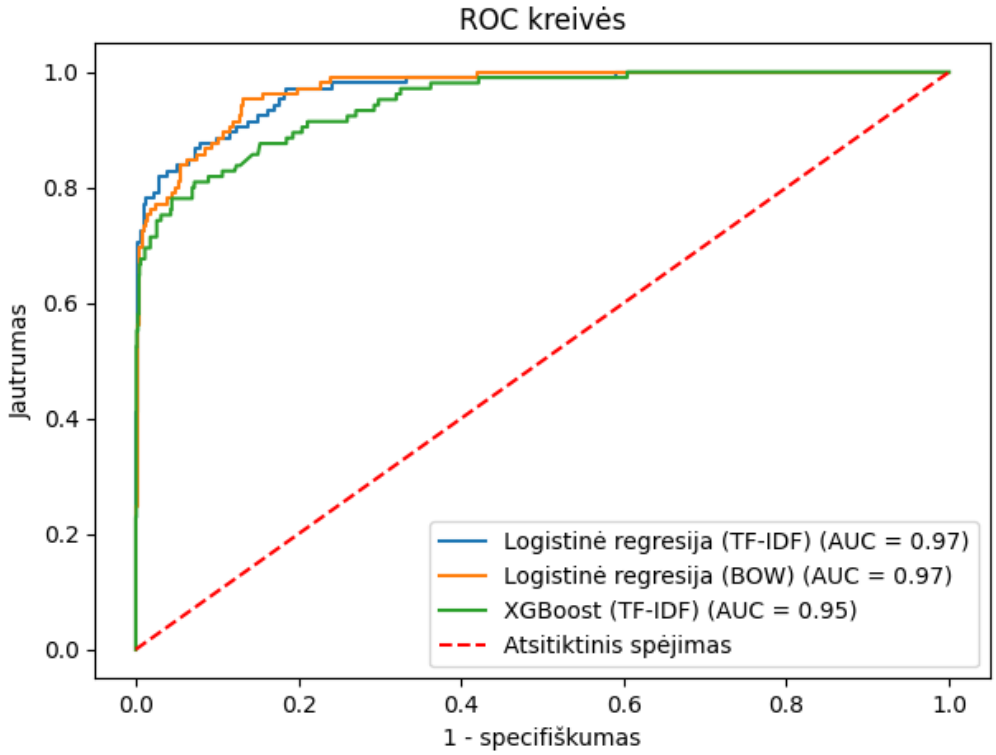
Modelis	**Bendras tikslumas**	**AUC**	**F1**	**Specifiškumas**	**Jautrumas**
Logistinė regresija (BOW)	0,973	0,971	0,882	0,852	0,714
Logistinė regresija (TF-IDF)	0,979	0,969	0,901	0,961	0,705
Logistinė regresija (SG)	0,973	0,944	0,869	0,918	0,638
Logistinė regresija (CBOW)	0,963	0,953	0,812	0,857	0,514
K-artimiausių kaimynų (BOW)	0,929	0,824	0,686	0,444	0,381
K-artimiausių kaimynų (TF-IDF)	0,973	0,945	0,863	0,969	0,600
K-artimiausių kaimynų (SG)	0,963	0,846	0,822	0,829	0,552
K-artimiausių kaimynų (CBOW)	0,949	0,805	0,747	0,662	0,429
Atsitiktinis miškas (BOW)	0,962	0,928	0,791	0,958	0,438
Atsitiktinis miškas (TF-IDF)	0,965	0,927	0,814	0,962	0,486
Atsitiktinis miškas (SG)	0,971	0,930	0,852	0,968	0,571
Atsitiktinis miškas (CBOW)	0,955	0,910	0,760	0,811	0,410
XGBoost (BOW)	0,968	0,952	0,849	0,865	0,610
XGBoost (TF-IDF)	0,973	0,948	0,869	0,918	0,638
XGBoost (SG)	0,970	0,933	0,855	0,878	0,619
XGBoost (CBOW)	0,955	0,913	0,769	0,780	0,438
Gradiento didinimas (BOW)	0,960	0,933	0,803	0,806	0,514
Gradiento dindinimas (TF-IDF)	0,964	0,893	0,828	0,822	0,571
Gradiento dindinimas (SG)	0,966	0,942	0,839	0,829	0,600
Gradiento dindinimas (CBOW)	0,949	0,917	0,740	0,689	0,400

Geriausio modelio sumaišymo matricoje (3.14 pav.) matome, kad klaidingi atsiliepimai buvo klasifikuoti labai tiksliai. Su teigiamais atsiliepimais situacija šiek tiek prastesnė, tačiau įvertinus klasių disbalansą, rezultatas tikrai tenkina išsikeltus lūkesčius.



3.14 pav. Geriausio modelio sumaišymo matrica

Nubraižius trijų geriausių modelių ROC kreives (3.15 pav.) galima nesunkiai pasakyti, kad geriausias klasifikavimo rezultatas yra logistinės regresijos modelių, nes jų kreivės yra arčiausiai viršutinio kairiojo kampo. Tuo tarpu raudona punktyrinė linija vaizduoja atsitiktinio spėjimo rezultatą.

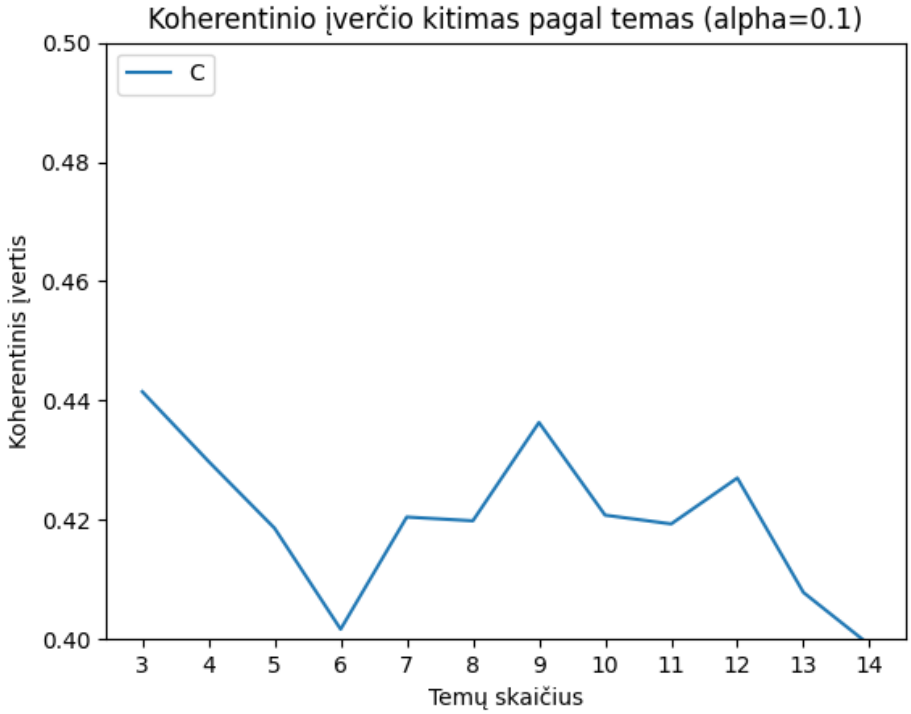


3.15 pav. Trijų geriausių modelių ROC kreivės

3.5. Temos modeliavimas naudojant neigiamus klientų atsiliepimus

Atliekant temos modeliavimą, vienas iš svarbiausių veiksmų yra tinkamas temų skaičiaus pasirinkimas. Temų skaičių nuspręsta pasirinkti atsižvelgiant į paskaičiuotą koherentinį įvertį. Kuo koherentinis įvertis didesnis, tuo temos geriau atsiskiria viena nuo kitos. Atliekant modeliavimą *alpha* parametras nebuvo derinamas ir parinkta statinė reikšmė – 0,1. Toliau esančiame 3.16 paveiksle

pateikiamas koherentinio įverčio kitimas paduodant skirtingą temų skaičių. Geriausias rezultatas gautas naudojant 3 ir 9 temas. Kadangi 3 temų koherentinis įvertis yra šiek tiek didesnis ir pastebėta, kad temose dominuojantys žodžiai yra logiškesni negu naudojant 9 temas, todėl nuspręsta naudoti būtent 3 temų LDA modelį.



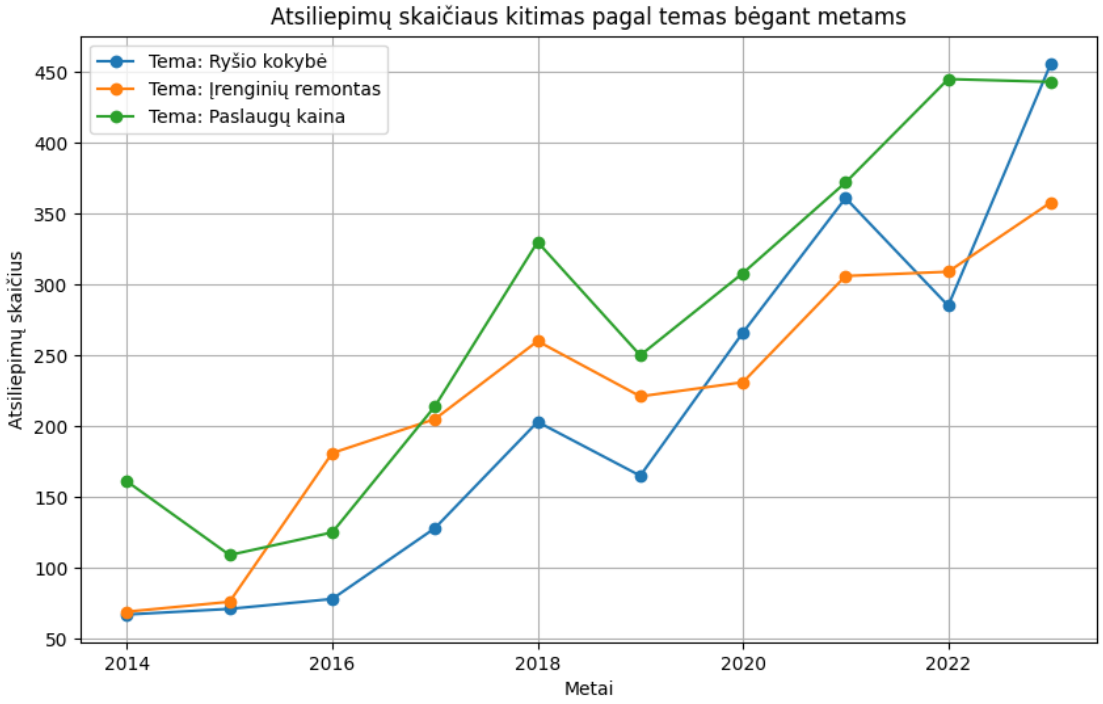
3.16 pav. Koherentinio įverčio kitimo pagal temas grafikas

Sukūrus 3 temų LDA modelį, kiekvienam atsiliepimui priskirta tema bei tikimybė, kuri nurodo atsiliepimo sąryšį su tema. Kuo tikimybė didesnė, tuo atsiliepimas geriau atspindi temą. Atsižvelgus į kiekvienoje temoje dominuojančius žodžius, sugalvoti tokie temų pavadinimai: ryšio kokybė, įrenginių remontas ir paslaugų kaina. Toliau pateiktoje 3.6 lentelėje pateikiami atsiliepimai, geriausiai reprezentuojantys gautas temas.

3.6 lentelė. Temas reprezentuojantys atsiliepimai

Tema	**Temą reprezentuojantys žodžiai**	**Atsiliepimas**
Ryšio kokybė	Ryšys, internetas, modemas, strigti, greitis	Esame TELIA vartotojai daugelį metų. Kol viskas veikė ir nereikėjo kreiptis dėl problemų, buvome patenkinti paslaugos kokybe, gaila, kad susidūrus su ja ir atsirado bloga aptarnavimo kokybė. Šį savaitgalį **likome be interneto ryšio ir televizijos paslaugų**...
Įrenginių remontas	Telefonas, draudimas, ekranas, garantija, remontas	Jau galvojau, kad mano situacija su Bite nueis iki teismo. Šiaip ne taip buvo išspręsta situacija, bet tai kainavo pinigus, daug nervų ir laiko. 2023 02 28 sudariau pokalbių plano sutratį, kad įsigyti telefona. **Telefonas** buvo įsigytas internetu su **atidaryta arba pažeista pakuote**...
Paslaugų kaina	Planas, mokestis, kaina, tinklas, sąskaita	Ilgai buvau patenkinta TELE2 Pildyk paslaugomis, ypač tenkino planai ir pasiūlymai, kur už **nedidelį mokestį** galima užsisakyti neribotus skambučius bei žinutes į visus tinklus. Deja, 2014 m. gruodžio 23d. užsisakiau planą VISIEMS galvodama, kad jis suteikia neribotus skambučius į visus tinklus, bet po poros savaičių naudojimosi planu nebegalėjau skambinti man staiga pasibaigė **sąskaita**...

Kadangi visiems atsiliepimams buvo priskirta viena iš 3 galimų temų, todėl nuspręsta paanalizuoti, kaip kito temų populiarumas bėgant metams (3.17 pav.). Žvelgiant į paskutinių kelių metų tendenciją, pastebima, kad mažiausiai atsiliepimų yra susijusių su įrenginių remontu. Tikėtina, kad klientai pastaraisiais metais rečiau naudojasi šia paslauga arba išspręstos anksčiau buvusios problemos. Ryšio kokybės tema analizuojamo laikotarpio pradžioje buvo retai pasitaikanti, tačiau per paskutinius 4 metus pastebimas ženklus augimas. Tam įtakos galėjo turėti COVID-19 pandemijos pradžia 2020 metais, kuomet klientai gerokai aktyviau pradėjo naudotis mobiliosiomis paslaugomis. Tuo tarpu kalbant apie paskutiniąją temą – paslaugų kaina, galima drąsiai teigti, kad tai yra klientams labai svarbi tema, kurios aktualumas per dešimtmetį ne tik nemažėjo, bet pastebimas gana ženklus pastarųjų keliolikos metų augimas.



3.17 pav. Atsiliepimų skaičiaus kitimo pagal temas bėgant metams grafikas

Sėkmingai išskyrus 3 dominuojančias temas, galima pereiti prie paskutinio tyrimų poskyrio, kuriame plačiau analizuojamas neigiamų atsiliepimų kontekstas.

3.6. Neigiamų atsiliepimų konteksto analizė

Ankstesniuose poskyriuose minėtas Word2Vec (skip-gram) metodas gali būti panaudotas ne tik vektorizavime, tačiau ir konteksto analizėje. Juo naudojantis galima nustatyti, koks kontekstas dažniausiai minimas šalia konkretaus žodžio. Šiame tyrime nuspręsta sukurti modelį, kuris vertina aplinkinius žodžius, nutolusius nuo tikslinio ne daugiau kaip per 10 pozicijų. Modelio kūrime pasirinktas žodžių vektoriaus dydis – 100.

Prieš naudojant modelį verta patikrinti jo veikimą. Pradžiai nuspręsta panaudoti kosinuso panašumo rodiklį, kuris leidžia palyginti panašius ir skirtingus žodžius. Modeliui paduoti 2 dažnai kartu vartojami žodžiai, kuriems kosinuso panašumas turėtų būti stiprus, o kiti 2 labai skirtingi ir vienas šalia kito sutinkami retai. Kaip ir buvo tikimasi, žodžiai "darbuotojas ir salonas" turi gerokai didesnį kosinuso panašumą negu „darbuotojas“ ir „savitarna“ (3.18 pav.).

```
Panašumas tarp žodžio "darbuotojas" ir "salonas":  0.67
Panašumas tarp žodžio "darbuotojas" ir "savitarna":  0.15
```

3.18 pav. Word2Vec (skip-gram) modelio testavimas, skaičiuojant kosinuso panašumą

Kadangi modelis neblogai susidorojo su ankstesne užduotimi, todėl nuspręsta patikrinti kitų žodžių panašumą. Pasirinkti 3 žodžiai, geriausiai reprezentuojantys ankstesniame poskyryje rastas temas. Gauti rezultatai (3.19 pav.) rodo, kad dažniausiai vienas šalia kito vartojami žodžiai „kaina“ ir „remontas“. Gana logiškas rezultatas, nes įrenginį remontuojančiam žmogui svarbi paslaugos kaina. Tuo tarpu mažiausias kosinuso panašumas rastas tarp žodžių „ryšys“ ir „remontas“.

```
Panašumas tarp žodžio "ryšys" ir "kaina":          0.31
Panašumas tarp žodžio "ryšys" ir "remontas":       0.26
Panašumas tarp žodžio "kaina" ir "remontas":       0.42
```

3.19 pav. Word2Vec (skip-gram) modelio rezultatai, skaičiuojant kosinuso panašumą

Antras būdas, kurį nuspręsta panaudoti modelio kokybės įvertinimui, yra nereikalingo žodžio radimas. Šiuo atveju modeliui nuspręsta paduoti įvairius įrenginių pavadinimus ir vieną žodį, kuris niekaip nesusijęs su įrenginiais. Modelis remdamasis žodžių panašumu turi rasti nereikalingą žodį ir jį grąžinti. Sukurtas modelis puikiai susidorojo su šia užduotimi ir grąžino nereikalingą žodį „darbuotojas“ (3.20 pav.).

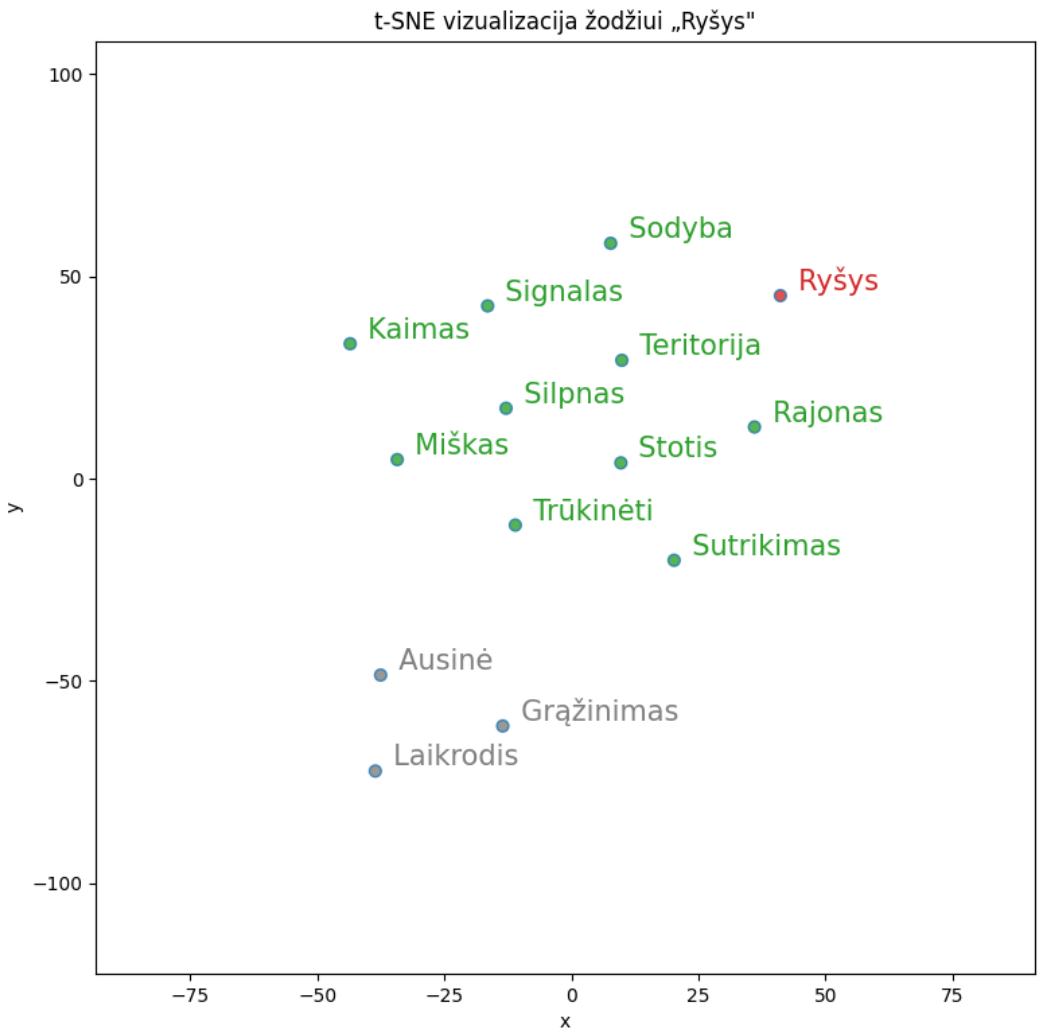
```
In [9]:
model_word2vec_sg.wv.doesnt_match([ 'kompiuteris', 'darbuotojas','telefonas',
'maršrutizatorius', 'planšetė', 'modemas'])

Out[9]:
'darbuotojas'
```

3.20 pav. Word2Vec (skip-gram) modelio testavimas, ieškant konteksto neatitinkančio žodžio

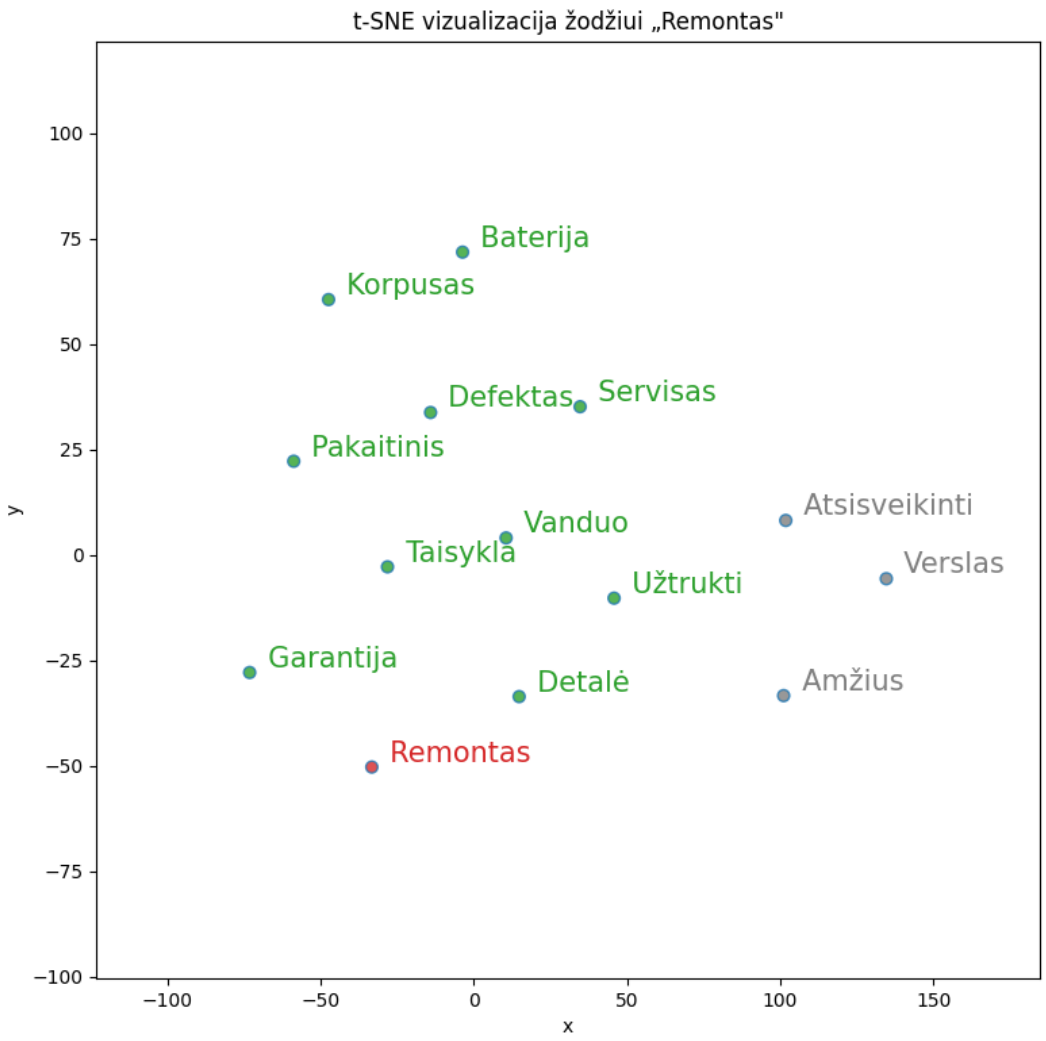
Praeitame poskyryje rastas 3 dominuojančias temas nuspręsta detaliau paanalizuoti pasiteikiant sukurta Word2Vec modelį. Kadangi Word2Vec modelis kiekvieną žodį saugo kaip vektorių, todėl nuspręsta sumažinti vektorių matmenis naudojant t-SNE (angl. *T-distributed stochastic neighbor embedding*) metodą ir tuomet vizualizuoti naudojant *matplotlib* biblioteką. Vizualizacijoje tikslinis žodis nuspalvintas raudona spalva, artimiausi žalia, o tolimiausi pilka spalva.

Pirmoji rasta dominuojanti tema susijusi su ryšio kokybe, todėl nuspręsta panaudoti žodį „ryšys“ ir rasti dažniausiai bei rečiausiai kontekste pasitaikančius žodžius. Toliau pateiktame 3.21 paveiksle galima pastebėti, kad dažnai kontekste minimi tokie žodžiai kaip „trūkinėti“ ir „silpnas“, kurie reiškia nestabilų ir kliento poreikių neatitinkantį ryšį. Taip pat verta atkreipti dėmesį, kad gana dažnai minimi žodžiai „sodyba“, „kaimas“ ir „miškas“, kas reiškia problemas su ryšiu atokesnėse Lietuvos vietovėse. Tuo tarpu žvilgtelėjus į tris rečiausius žodžius, minimus ryšio kontekste, galima pastebėti, kad jie susiję su įrenginiais bei jų grąžinimu. Tai reiškia, kad įrenginių gražinimo tema yra labiausiai nutolusi nuo ryšio ir klientai jas kartu naudoja labai retai.



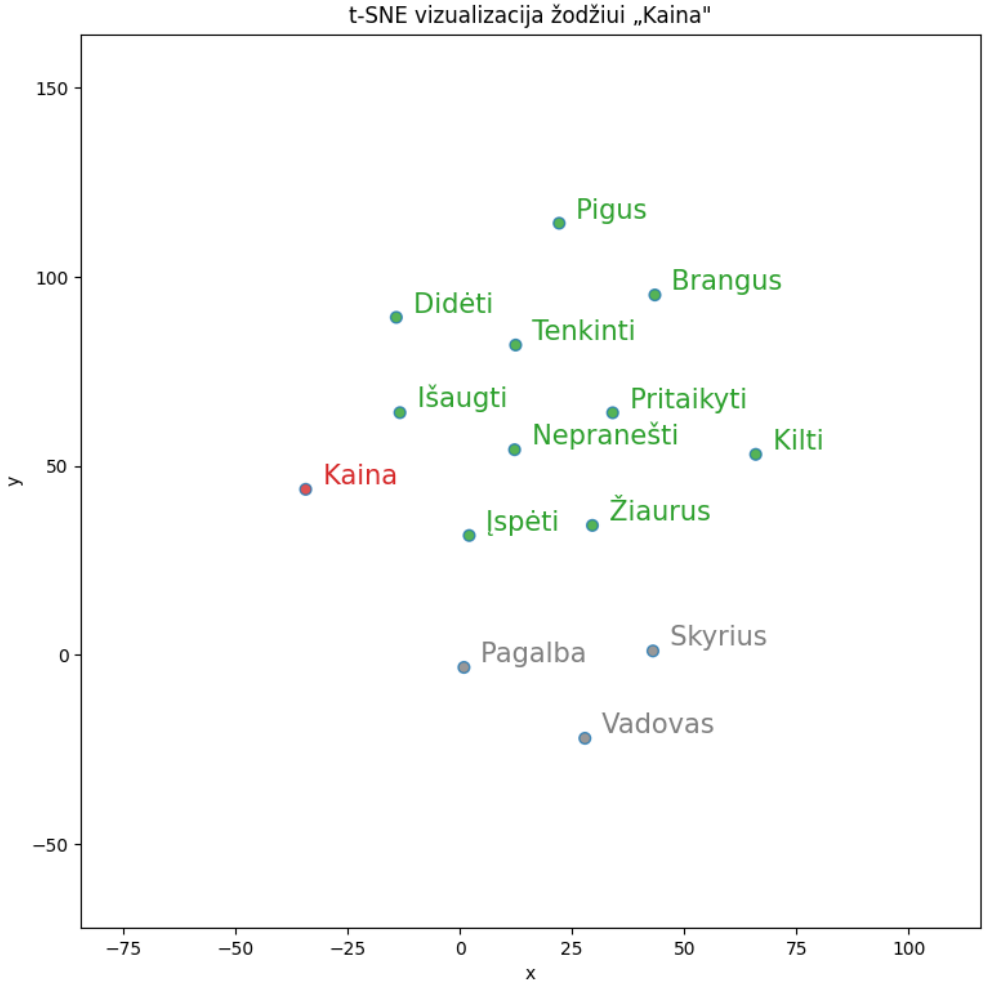
3.21 pav. t-SNE vizualizacija žodžiui „Ryšys“

Antroji tema susijusi su įrenginių remontu, todėl pasirinktas tikslinis žodis – „remontas“. Sudarytoje vizualizacijoje (3.22 pav.) galima pastebėti, kad kontekste dažniausiai minimi žodžiai, susiję su remontuojamo įrenginio dalimi: baterija, korpusas, stiklas. Be to, nemažai klientų susiduria su gamykliniais įrenginių defektais ir šias iškilusias problemas bando išspręsti pasinaudodami dar galiojančia gamintojo garantija. Antroji remonto priežastis – vanduo, dėl kurio padarytos žalos klientams tenka kreiptis į įrenginių taisyklą. Žodis „užtrukti“ sufleruoja, kad dažnai įrenginių remontas trunka ilgiau nei buvo tikimasi. Analizuojant su remontu mažiausiai susijusius žodžius, galima įžvelgti verslo kontekstą. Galima daryti prielaidą, kad verslo klientai kur kas rečiau naudojasi remonto paslaugomis lyginant su privačiais vartotojais.



3.22 pav. t-SNE vizualizacija žodžiui „Remontas“

Paskutinioji trečioji tema susijusi su paslaugų kaina, todėl naudojamas tikslinis žodis – „kaina“. Vizualizacijoje (3.23 pav.) pastebima, kad dažnai kontekste kalbama apie kainos didėjimą. Tai įrodo tokie žodžiai kaip „kilti“, „didėti“, „išaugti“. Taip pat dažnai minimi žodžiai „įspėti“ ir „nepranešti“. Galima daryti prielaidą, kad klientai apie didėjančią paslaugų kainą sužino tik iš gautos sąskaitos be jokios išankstinės žinios. Tuo tarpu kalbant apie labiausiai nuo kainos konteksto nutolusius žodžius, galima pastebėti, kad tai yra su aptarnavimu susiję žodžiai. Galima teigti, kad kaina besiskundžiantys klientai tuo pat metu labai retai užsimena apie aptarnavimo problemas.



3.23 pav. t-SNE vizualizacija žodžiui „Kaina“

Sukurtas Word2Vec modelis padėjo išsamiau paanalizuoti pasirinktų žodžių kontekstą. Pavyko rasti ne tik dažniausiai tikslinių žodžių kontekste pasitaikančius žodžius, tačiau ir tuos, kurie beveik niekada nenaudojami greta analizuotų žodžių.

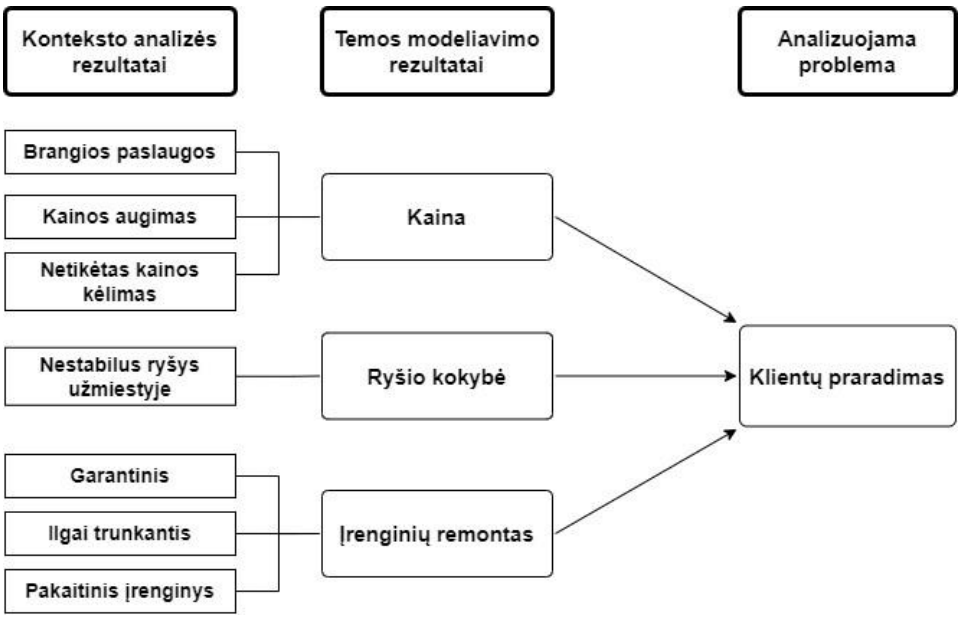
3.7. Tyrimo rezultatų apibendrinimas

Atlikus visus planuotus tyrimus, galima vėl grįžti prie konceptualiojo modelio ir palyginti teorinius klientų praradimą lemiančius veiksnius su praktiniais. Toliau pateiktoje 3.7 lentelėje išvardyti visi klientų praradimą lemiantys veiksniai, kurie buvo apibrėžti konceptualiajame modelyje bei gauti atlikus praktinį tyrimą. Galima pastebėti, kad 2 veiksniai buvo identifikuoti tiek teorinėje dalyje, tiek praktinėje. Galima drąsiai teigti, kad kaina bei ryšio kokybė yra svarbūs bei klientų praradimą lemiantys veiksniai. Teorinėje dalyje buvo pasirinkti papildomi 3 veiksniai, kurių praktinėje dalyje nepavyko išskirti: aptarnavimas, reklama bei savitarnos sistema. Tačiau nuspręsta patikrinti, koks kontekstas dominuoja šalia šių žodžių. Šalia žodžio „reklama“ dažnai yra minima melaginga informacija. Tai reiškia, kad klientai dažniausiai skundžiasi ne dėl jų pažiūrų neatitinkančios reklamos, bet dėl to, kad skelbiama ne visiškai teisinga informacija. Paskutinis veiksnys, kurį pavyko identifikuoti tik tyrimo metu – įrenginių remontas. Šio veiksnio identifikavimas reiškia, kad klientai dažnai kreipiasi į telekomunikacijų įmones dėl įrenginių gedimų ir šios paslaugos suteikimas neapsieina be papildomų problemų.

3.7 lentelė. Klientų praradimą lemiančių veiksnių konteksto analizės rezultatai

Veiksnys	Teorinis/praktinis	Tikslinis žodis	Dažniausi konteksto žodžiai	Rečiausi konteksto žodžiai
Kaina	Teorinis ir praktinis	Kaina	• Pigus • Didėti • Brangus	• Pagalba • Skyrius • Vadovas
Ryšio kokybė	Teorinis ir praktinis	Ryšys	• Trūkinėti • Teritorija • Signalas	• Laikrodis • Grąžinimas • Ausinė
Aptarnavimas	Teorinis	Aptarnavimas	• Bendravimas • Lankyti • Kultūra	• Tarnyba • Atsargus • Standartinis
Reklama	Teorinis	Reklama	• Melas • Melagingas • Investuoti	• Prašymas • Mokėtojas • Kodas
Savitarnos sistema	Teorinis	Savitarna	• Versija • Prisijungimas • Svetainė	• Ausinė • Servisas • Perkant
Įrenginių remontas	Praktinis	Remontas	• Garantija • Defektas • Vanduo	• Amžius • Verslas • Atsisveikinti

Šiame projekte atliktas temo modeliavimas leido identifikuoti 3 klientų praradimą telekomunikacijų paslaugų sektoriuje lemiančius veiksnius: kaina, ryšio kokybė bei įrenginių remontas (3.24 pav.). Konteksto analizė padėjo giliau paanalizuoti visus 3 temos modeliavimo metu rastus klientų praradimą lemiančius veiksnius. Remiantis gautais rezultatais, kainą būtų galima skelti į 3 atskiras šakas: brangios paslaugos, kainos augimas, netikėtas kainos kėlimas. Analizuojant ryšį pastebėta, kad klientai dažniausiai skundžiasi nestabiliu ryšiu užmiestyje. O įrenginių remontas išskirtas į tris atskiras šakas: garantinis, ilgai trunkantis ir pakaitinis įrenginys.



3.24 pav. Apibendrinti tyrimo rezultatai

Sekančiame skyriuje pateikiamos rekomendacijos, siekiant mažinti klientų praradimą telekomunikacijų paslaugų sektoriuje.

3.8. Siūlomi klientų praradimo mažinimo telekomunikacijų paslaugų sektoriuje sprendimai

1. **Lankstūs mokėjimo planai**. Tyrimo metu pastebėta, kad telekomunikacijų paslaugų sektoriuje nemaža dalis klientų skundžiasi aukštomis paslaugų kainomis. Vienas iš galimų šios problemos sprendimų yra tikslių klientų poreikių išsiaiškinimas ir lankstesnių planų, kurie geriausiai atitiktų klientų finansines galimybes bei poreikius, pasiūlymas. Šis sprendimas labiau orientuotas į paslėptą kliento praradimo tipą, kuomet klientas jau seniai aktyviai nesinaudoja paslaugomis arba naudoja tik labai mažą dalį plano teikiamų galimybių.
2. **Konkurencinga paslaugų kaina**. Norint sėkmingai konkuruoti rinkoje, būtina stebėti konkurentų siūlomas paslaugų kainas. Mažėjančios konkurentų siūlomos paslaugų kainos turi didelę įtaką klientų praradimui, todėl įmonei būtina ne tik neatsilikti nuo konkurentų, bet ir stengtis pasiūlyti geriausią kainos ir kokybės santykį.
3. **Tinklo infrastruktūros plėtra**. Atliktas tyrimas parodė, kad klientai dažnai skundžiasi nestabiliu mobiliuoju ryšiu užmiestyje. Norint spręsti šią problemą, būtina investuoti į tinklo infrastruktūros plėtrą, ypač mažiau aprūpintose vietovėse. Naujų technologinių sprendimų įgyvendinimas leistų pagerinti ryšio stabilumą ir klientams, gyvenantiems atokesnėse vietovėse, nereikėtų nerimauti dėl prasto ryšio bei svarstyti apie konkurentų siūlomų paslaugų pasirinkimą. Šis sprendimas orientuotas į dalinį kliento praradimo tipą, kuomet klientas naudojasi keliomis paslaugomis vienu metu ir renkasi, kurios yra kokybiškesnės ir labiau atitinkančios jo individualius poreikius.
4. **Įrenginio draudimo paslauga**. Tyrimo rezultatai atskleidė, kad daugelis klientų susiduria su neplanuotu įrenginio gedimu, atsiradusiu dėl jo dužimo ar vandens padarytos žalos. Įrenginio draudimo paslauga klientui leistų apsisaugoti nuo netikėtų išlaidų, o mobiliųjų paslaugų teikėjui tai leistų stiprinti kliento lojalumą.

Išvados

1. Atlikta mokslinės literatūros analizė leidžia teigti, kad klientų praradimas dažniausiai apibrėžiamas kaip situacija, kai kliento indėlis į įmonės pajamas mažėja. Klientų praradimas dažnai skirstomas į du stambius tipus: su sutartimi ir be sutarties. Šie du tipai skirstomi į dar smulkesnius, kurie leidžia ne tik geriau pažinti klientą, bet ir imtis reikiamų veiksmų siekiant jo neprarasti.
2. Išanalizavus iki šiol atliktų mokslinių tyrimų rezultatais pagrįstus klientų praradimo valdymo modelius ir jų taikymo telekomunikacijų paslaugų sektoriuje galimybes, sudarytas konceptualusis klientų praradimą telekomunikacijų paslaugų sektoriuje lemiančių veiksnių modelis. Į šį modelį įtraukti penki kritiškai svarbiais laikomi klientų praradimą lemiantys veiksniai – kaina, ryšio kokybė, aptarnavimas, reklama ir savitarnos sistema.
3. Klientų atsiliepimų, paskelbtų skaitmeninėse platformose, klasifikavimui pagal sentimentus pasirinkti 5 mašininio mokymosi ir 4 vektorizavimo metodai, kurie bene dažniausiai naudojami įvairiuose tyrimuose. Tuo tarpu neigiamų atsiliepimų temos modeliavimui pasirinktas LDA metodas, kurio rezultatai pateikti ne viename analizuotame natūralios kalbos apdorojimo tyrime. Kadangi Word2Vec metodas naudotas vektorizavimo etape, todėl nuspręsta išbandyti jo kitą siūlomą galimybę – konteksto analizė.
4. Iš pasirinktų 5 skaitmeninių platformų, daugiausiai klientų atsiliepimų rasta Rekvizitai.lt svetainėje. Analizuojant klientų atsiliepimų sentimentus pastebėta, kad teigiami atsiliepimai sudaro tik maždaug 6 % visų atsiliepimų. Klasifikuojant sentimentus geriausias rezultatas gautas naudojant logistinę regresiją bei TF-IDF vektorizavimo metodą. Šio modelio F1 įvertis siekia 90,1 %.
5. Temos modeliavimas padėjo išskirti 3 neigiamuose klientų atsiliepimuose dominuojančias temas, lemiančias klientų praradimą telekomunikacijų paslaugų sektoriuje. Tyrimo metu rastos temos: kaina, ryšio kokybė ir įrenginių remontas. Konteksto analizės etape atidžiau pasigilinta į temos modeliavime išskirtas temas. Pastebėta, kad klientai dažnai skundžiasi nestabiliu mobiliuoju ryšiu užmiestyje, didelėmis ir nuolat didėjančiomis paslaugų kainomis bei ilgai trunkančiu įrenginių remontu.
6. Tyrimo metu identifikuoti 2 veiksniai, kurie taip pat buvo pateikti ir konceptualiajame modelyje. Galima teigti, kad kaina ir ryšio kokybė neabejotinai lemia klientų praradimą telekomunikacijų paslaugų sektoriuje. Trečiasis veiksnys, kurio nebuvo konceptualiajame modelyje, bet jis identifikuotas praktinėje dalyje, tai įrenginių remontas. Remiantis gautais tyrimo rezultatais, suformuluotos 4 rekomendacijos, siekiant mažinti klientų praradimą telekomunikacijų paslaugų sektoriuje: didesnė lanksčių mokėjimo planų pasiūla, konkurencingos kainos pateikimas, tinklo infrastruktūros modernizavimas ir įrenginių draudimo paslaugos tobulinimas.

Literatūros sąrašas

1. Sathish, M., Santhosh, K.K., Naveen, K.J. and Jeevanantham, V. (2011). A study on consumer switching behaviour in cellular service provider: a study with reference to Chennai, Far East Journal of Psychology and Business, Vol. 2 No. 2, p. 71-81.
2. Amoako, G., Arthur, E., Bandoh, C. and Katah, R. (2012). The impact of effective customer relationship management (CRM) on repurchase: a case study of (GOLDEN TULIP) hotel (ACCRA-Ghana), Journal of Marketing Management, Vol. 4 No. 1, p. 7-29.
3. Ascarza, E. (2018), "Retention futility: targeting high-risk customers might be ineffective", Journal of Marketing Research, Vol. 55 No. 1, p. 80-98.
4. Klimavičienė, K., Lingaitienė O. (2019). Ryšių su klientais valdymo (angl. customer relationship management) sistemų diegimo problematika Lietuvoje. http://jmk.vvf.vgtu.lt/index.php/Verslas/2019/paper/viewFile/380/153
5. Keropyan, A., Gil-Lafuente, A. M. (2012). Customer loyalty programs to sustain consumer fidelity in mobile telecommunication market. Expert Systems with Applications Vol. 39 No. 12, p. 11269-11275.
6. Darzi, M. A., Bhat, S. A. (2018). Personnel capability and customer satisfaction as predictors of customer retention in the banking sector: A mediated-moderation study. International Journal of Bank Marketing, Vol 36 No. 4, p. 663-679.
7. Huarng, K., Hui-Kuang Yu, T. (2020). The impact of surge pricing on customer retention. Journal of Business Research, Vol. 120, p. 175-180.
8. Mahmoud, M. A., Hinson, R. E., Adika, M. K. (2018). The Effect of Trust, Commitment, and Conflict Handling on Customer Retention: The Mediating Role of Customer Satisfaction. Journal of Relationship Marketing, Vol. 17 No. 4, p. 257-276. https://doi.org/10.1080/15332667.2018.1440146
9. Seo, D., Ranganathan, S. and Yair, B. (2008). Two-level model of customer retention in the US mobile telecommunications service market, Telecommunications Policy, Vol. 32 Nos 3/4, p. 182-196.
10. Shobana, J., Gangadhar, C., Arora R. K., Renjith P.N., Bamini J., Chincholkar Y. D. (2023). E-commerce customer churn prevention using machine learning-based business intelligence strategy. Measurement: Sensors, Vol. 27.
11. Skačkauskienė, I., Toropovaitė, K. (2011). Ryšių marketingo kaip vartotojų lojalumą formuojančio veiksnio tyrimas. Contemporary Issues in Business, Management and Education 2011, p. 264-276.
12. Lazarov, V., Capota, M.(2007). Churn Prediction. Bus. Anal. Course TUM Comput. Sci.
13. Xu,T., Ma, Y., Kim, K. (2021). Telecom Churn Prediction System Based on Ensemble Learning Using Feature Grouping. Applied Sciences, Vol 11 No. 11.
14. Lappeman, J., Franco, M., Warner V., Sierra-Rubia, L. (2022). What social media sentiment tells us about why customers churn. Journal of Consumer Marketing, Vol. 39 No. 5, p. 385-403.
15. Lee, E., Kim, J., Lee, S. (2017). Predicting customer churn in mobile industry using data mining technology. Industrial Management & Data Systems, Vol. 117 No. 1, p. 90-109.
16. Bhattacharyya, J., Dash, M. K. (2021). Investigation of customer churn insights and intelligence from social media: a netnographic research. Online Information Review, Vol. 45 No. 1, p. 174-206.

17. Lamrhari, S., Ghazi, H. E., Oubrich, M., Faker, A. E. (2022). A social CRM analytic framework for improving customer retention, acquisition, and conversion. Technological Forecasting and Social Change, Vol. 174.
18. Hejazinia, R. and Kazemi, M. (2014). Prioritizing factors influencing customer churn. Interdisciplinary Journal of Contemporary Research in Business, Vol. 5 No. 12, p. 227-236. https://journal-archieves36.webs.com/227-236apr14.pdf
19. Lima Lemos, R. A., Silva, T. C., Tabak, B. M. (2022). Propension to customer churn in a financial institution: a machine learning approach. Neural Computing and Applications, Vol. 34, p. 11751–11768.
20. Trivedi, S. K., Singh, A. (2021). Twitter sentiment analysis of app based online food delivery companies. Global Knowledge, Memory and Communication, Vol. 70 No. 8/9, p. 891-910.
21. He, W., Tian, X., Wang, F. (2019). Innovating the customer loyalty program with social media: A case study of best practices using analytics tools. Journal of Enterprise Information Management, Vol. 32 No. 5, p. 807-823.
22. Aljedaani, W., Rustam, F., Mkaouer, M. W., Mkaouer, M. W., Ghallab, A., Rupapara, V. Washington, P. B., Lee, E., Ashraf, I. (2022). Sentiment analysis on Twitter data integrating TextBlob and deep learning models: The case of US airline industry. Knowledge-Based Systems, Vol. 255.
23. Suhaimin, M. S. M., Hijazi, M. H. A., Moung, E. G., Nohuddin, P. N. E., Chua, S., Coenen, F. (2023). Social media sentiment analysis and opinion mining in public security: Taxonomy, trend analysis, issues and future directions. Journal of King Saud University - Computer and Information Sciences, Vol. 35.
24. Sidya, N. A., Fanany, M. I., Budi, I. (2015). Twitter Sentiment to Analyze Net Brand Reputation of Mobile Phone Providers. Procedia Computer Science, Vol. 72, p. 519-526.
25. Qamar, A. M., Sohali, S. (2017). Sentiment Classification of Twitter Data Belonging to Saudi Arabian Telecommunication Companies. International Journal of Advanced Computer Science and Applications, Vol. 8 No. 1, p. 395-401.
26. Biradar, S. H., Gorabal, J.V., Gupta G. (2022). Machine learning tool for exploring sentiment analysis on twitter data. Materials Today: Proceedings, Vol. 56 No. 4, p. 1927-1934.
27. Štrimaitis, R., Stefanovič, P., Ramanauskaitė, S., Slotkienė., A. (2021). Financial Context News Sentiment Analysis for the Lithuanian Language.
28. Kapočiūtė-Dzikienė, J., Salimbajevs, A. (2022). Comparison of Deep Learning Approaches for Lithuanian Sentiment Analysis. Baltic J. Modern Computing, Vol. 10 No. 3, p. 283–294.
29. Diekson, Z. A., Bagas Prakoso, M. R., Qalby Putra, M. S., Al Fadel Syaputra, M. S., Achmad, S., Sutoyo, R. (2022). Sentiment analysis for customer review: Case study of Traveloka. Procedia Computer Science, Vol. 216, p. 682–690.
30. Dzisevič, R., Šešok, D. (2019). Text Classification using Different Feature Extraction Approaches.
31. Umarania, V., Juliana, A., Deepa, J. (2021). Sentiment Analysis using various Machine Learning and DeepLearning Techniques. Journal of the Nigerian Society of Physical Sciences, Vol. 3, p. 185-394.
32. AminiMotlagh, M., Shahhoseini, H., Fatehi, N. (2023). A reliable sentiment analysis for classification of tweets in social networks. Social Network Analysis and Mining, Vol. 13 No. 7. https://doi.org/10.1007/s13278-022-00998-2.

33. Kapočiūtė-Dzikienė, J., Damaševičius, R., Woźniak, M. (2019). Sentiment Analysis of Lithuanian Texts Using Traditional and Deep Learning Approaches. Computers Vol. 8 No. 1. https://doi.org/10.3390/computers8010004.
34. Abou el Kassem, E., Ali Hussein, S., Mostafa Abdelrahman, A., Kamal Alsheref, F. (2020). Customer Churn Prediction Model and Identifying Features to Increase Customer Retention based on User Generated Content. International Journal of Advanced Computer Science and Applications, Vol. 11 No. 5, p. 522-531.
35. Jeelall, S., Cheerkoot-Jalim, S. (2020). HealthMine: A Tool for Social Media Text Mining in Health. 2020 3rd International Conference on Emerging Trends in Electrical, Electronic and Communications Engineering (ELECOM), p. 53-57.
36. Bikku, T., Jarugula, J., Kongala, L., Tummala, N. D., Donthiboina, N. V. (2023). Exploring the Effectiveness of BERT for Sentiment Analysis on Large-Scale Social Media Data. 2023 3rd International Conference on Intelligent Technologies (CONIT), p. 1-4.
37. Lubis, A. R., Fatmi, Y., Witarsyah, D. (2023). Comparison of Transformer Based and Traditional Models on Sentiment Analysis on Social Media Datasets. 2023 6th International Conference of Computer and Informatics Engineering (IC2IE), p. 163-168.
38. Singh, A., Srivastava, H., Aman, M., Dubey, G. (2023). Sentiment Analysis on User Feedback of a Social Media Platform. 2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), p. 826-832.
39. Bhattacharjee, D., Paul, A., Kumar, D. (2023). Sentiment Analysis for Hateful Content on Social Media. 2023 International Conference on Network, Multimedia and Information Technology (NMITCON), p. 1-6.
40. Mahmud, S., Islam, T., Jaman Bonny, A., Khatun Shorna, R., Hossain Omi, J., Rahman, S. (2022). 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT), p. 1-6.
41. Taherkhani, L., Daneshvar, A., Khalili, H. A., Sanaei, M. R. (2023). Analysis of the Customer Churn Prediction Project in the Hotel Industry Based on Text Mining and the Random Forest Algorithm. Advances in Civil Engineering, Vol. 2023. https://doi.org/10.1155/2023/6029121
42. Alshamari, M. A. (2023). Evaluating User Satisfaction Using Deep-Learning-Based Sentiment Analysis for Social Media Data in Saudi Arabia's Telecommunication Sector. Computers 2023, Vol. 12, No. 170. https://doi.org/10.3390/computers12090170
43. Wayasti, R. A., Surjandari, Zulkamain, I. (2018), Mining Customer Opinion for Topic Modeling Purpose: Case Study of Ride-Hailing Service Provider. 2018 6th International Conference on Information and Communication Technology (ICoICT), Bandung, Indonesia, p. 305-309.
44. Yamunathangam, D., Bharathi Priya, C., Shobana, G., Latha, L. (2021). An Overview of Topic Representation and Topic Modelling Methods for Short Texts and Long Corpus. 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), Coimbatore, India, p. 1-6.