Draudimo sektoriaus klientų atsiliepimų ir vertinimų nuotaikų kaitos analizė laike

Donata Petkutė, Gražina Korvel

Vilniaus universitetas, Matematikos ir informatikos fakultetas,
Duomenų mokslo ir skaitmeninių technologijų institutas,
Akademijos g. 4, Vilnius
donata.petkute@mif.stud.vu.lt

Santrauka. Šiandien internetas tampa nepakeičiamas informacijos šaltinis, kuriame gausu įvairių atsiliepimų apie įsigytus produktus ar paslaugas. Šie atsiliepimai teikia vertingą informaciją įmonėms, norinčioms geriau suprasti savo klientų poreikius ir lūkesčius. Vienas iš efektyviausių būdų išgauti įžvalgas iš atsiliepimų yra naudoti nuotaikų analizę. Šiame tyrime aptariama, kiek klientų yra patenkinti ir nepatenkinti draudimo sektoriaus teikiamomis paslaugomis bei siūlomais produktais, taip pat nuotaikų priskyrimui buvo naudojami skirtingi vektorizavimo ir klasifikavimo metodai, kad būtų pasiekti geriausi rezultatai. Analizei atlikti naudojami du duomenų rinkiniai – produkto įsigijimo ir atsiliepimai apie žalos išmokėjimą po draudiminio įvykio. Tyrime naudojami du vektorizavimo būdai – žodžių maišo ir TF-IDF bei trys klasifikavimo metodai: atraminių vektorių, naiviojo Bajeso bei ilgalaikės trumposios atminties modelis. Atlikus tyrimą gauta, jog klientų atsiliepimų nuotaikas geriausiai klasifikuoja naiviojo Bajeso klasifikatorius su TF-IDF vektorizavimo būdu, kai tikslumas siekia 91% abiem duomenų rinkiniams. Atsiliepimams po produkto įsigijimo gautos preciziškumo ir atkūrimo metrikos teigiamam sentimentui 93% ir 97% atitinkamai, neigiamai klasei 73% ir 55%. Teigiamai klasei po žalų atlyginimo gautas preciziškumas 93% ir atkūrimo metrika 96%, o neigiamai – 82% ir 72% . Pritaikius atraminių vektorių klasifikatorių su skirtingomis vektorizavimo technikomis gauta tikslumo įvertis 89%.

Raktiniai žodžiai: nuotaikų analizė, natūralios kalbos apdorojimas, mašininis mokymasis, TF-IDF, žodžių maišas.

1 Įvadas

Šiais laikais socialinė žiniasklaida atlieka labai svarbų vaidmenį beveik kiekvieno žmogaus kasdieniame gyvenime, pavyzdžiui, suteikia vartotojams galimybę išreikšti savo nuomonę apie tam tikrą produktą ar paslaugą [6, 10]. Iš tikrųjų vis dažniau žmonės pasikliauna kitų klientų patirtimi ir naudojasi atsiliepimų apžvalgomis, kurios padeda nuspręsti produkto įsigijimo svar-

DOI: https://doi.org/10.15388/LMITT.2024.15

bą. Vieni žmonės produktui skiria keturis ar penkis balus ir išreiškia galutinį pasitenkinimą produktu, o kiti skiria vieną ar du − išreikšdami visišką nepasitenkinimą. Tai nekelia jokių sunkumų siekiant suprasti klientų nuotaiką. Tačiau kiti žmonės skiria tris balus, nors akivaizdžiai išreiškia visišką pasitenkinimą produktu. Tai klaidina kitus klientus, taip pat įmones, norinčias sužinoti tikrąją nuomonę [2]. Įmonėms tampa būtina suprasti žmonių nuotaikas, tam kad galėtų ištirti vartotojų nuomonę ir požiūrį į jų paslaugas bei atrastų naujų verslo strategijų [3, 12]. Siekiant geriau suvokti bendrą vartotojų požiūrį ir pasitenkinimą teikiamomis paslaugomis, verta atlikti nuotaikų analizę, kurios metu bandoma nustatyti ar sakinys yra teigiamas, ar neigiamas. [10, 11]. Dažnai produktai neatitinka klientų lūkesčių, todėl nuotaikų analizės naudojimas po naujo produkto išleidimo į rinką, bendrovėms gali padėti suprasti jo trūkumus ir pranašumus [3]. Tačiau visuomenės nuomonė apie tam tikrą temą laikui bėgant kinta, todėl siekiant nustatyti tendencijas bei sezoniškumą svarbu atlikti analizę laike. Be to, tokia analizė padeda nustatyti nukrypimus, kurie gali būti susiję su įvykiais, sukėlusiais nuotaikų pokyčius [4].

M. F. Madjid ir kt. atliko nuotaikos analizę, nagrinėdami programų atsiliepimus, naudodami atraminių vektorių modelį (angl. Support vector machine, SVM) ir naiviojo Bajeso (angl. Naive Bayes, NB) klasifikatorių [16]. Tyrime gauti rezultatai rodo, kad atraminių vektorių metodo tikslumas pasiekia 94,29 %, o Naiviojo Bajeso klasifikatorius – 93,97 % tikslumą. Tuomet atliekant oro linijų atsiliepimų nuotaikų analizę A. M. Rohat ir kt. naudojo naiviojo Bajeso ir atraminių vektorių modelį. Rezultatai taip pat parodė, kad oro linijų atsiliepimų atveju SVM užtikrina daug geresnius tikslumo rezultatus (82 %), o NB algoritmas – tik 76 % [14]. Kitą sentimentų klasifikavimo analizę atliko J.J. A. Limbong ir kiti tyrėjai, kuriuo metu buvo naudojami naiviojo Bajeso ir k artimiausių kaimynų (angl. k-nearest neighbors, KNN) klasifikatoriai, rezultatai parodė, kad KNN metodas veikia geriau, jo klasifikavimo tikslumo vertė yra 92,8 %, palyginti su NB metodo tikslumo verte – 91,4 % [15]. Taip pat pastaraisiais metais gilusis mokymasis sulaukia vis daugiau dėmesio pramonėje ir akademiniame pasaulyje dėl savo didelio našumo įvairiose srityse. Šiuo metu populiariausi giliojo mokymosi architektūros tipai yra pasikartojantis neuroninis tinklas (angl. Reccurent Neural Network, RNN) ir konvoliucinis neuroninis tinklas (angl. Convolutional Neural Network, CNN) [7]. Straipsnyje [11] tyrėjai taikė gilaus mokymosi metodus − konvoliucinį neuroninį tinklą, ilgosios trumpalaikės atminties (angl. Long Short-Term

Memory, LSTM) modelį bei paprastąjį neuroninį tinklą (angl. Simple Neural Network), kad atliktų Twitter nuotaikų analizę. Autoriai gavo, jog LSTM yra geriausias iš visų naudotų metodų, jo tikslumas yra 87 %, o CNN ir paprastojo neuroninio tinklo metodai atitinkamai pasiekė 82 % ir 81 % tikslumą. Kiekvienu atveju norint įvertinti klasifikavimo rezultatus buvo naudojamos tikslumo metrikos: tikslumas, preciziškumas, atkūrimas, F1 statistika.

2 Pirminis teksto apdorojimas

Pirminis teksto apdorojimas taikomas siekiant išvalyti ir paruošti tekstą nuotaikų klasifikavimui, nes dažnu atveju vartotojų rašomi tekstai yra nestruktūrizuoti. Tokiuose tekstuose paprastai yra daug nereikalingos, nenaudingos informacijos, pavyzdžiui, pasikartojančių žodžių, skaičių, skyrybos ženklų, rašybos klaidų, jaustukų ir sutrumpinimų. Darbe pirmiausia atsiliepimai buvo suskaidyti į teksto vienetus. Toliau pašalinami nereikšmingi žodžiai, tai gali būti jungtukai, įvardžiai, kurie laikomi nereikalingais ir nenaudingais.

3 Teksto vektorizavimas

Atlikus pirminį teksto apdorojimą, kitas žingsnis pritaikyti vektorizavimą, kuris tekstinius duomenis paverčia skaitiniu vektoriumi. Darbe taikomi du skirtingi vektorizavimo būdai – žodžių maišo ir TF-IDF. Kadangi dauguma mašininio mokymosi algoritmų duomenis apdoroja su skaitiniais įvesties reikšmėmis, tai yra būtinas žingsnis atliekant nuotaikų analizę.

Žodžių maišas (angl. Bag-of-words, BOW) − tai natūraliosios kalbos apdorojimo metodas, naudojamas tekstiniam dokumentui atvaizduoti kaip žodžių rinkiniui, neatsižvelgiant į jų pateikimo tvarką. Žodžių maišo metodas yra vienas paprasčiausių tokio tipo metodų, skaičiavimo bei konceptualumo prasme. Pagrindinė idėja – suskaičiuoti kiekvieno žodžio dažnumą dokumente ir remiantis šiais žodžių dažniais, sukurti dokumento vektorinį atvaizdavimą [30].

Termino dažnis-atvirkštinio dokumento dažnis (angl. Term Frequency-Inverse Document Frequency, TF-IDF) populiarus tyrimų metodas natūralios kalbos apdorojimo srityje. TF-IDF metodu nustatomas santykinis žodžių dažnis konkrečiame dokumente. Žodžiai, kurie tekste pasitaiko dažniau yra laikomi mažiau svarbiais, o rečiau pasitaikantys žodžiai priskiriami svarbesniems, nes laikoma, kad jie turi daugiau reikšmingos informacijos [1]. Metodo matematinė išraiška parodyta žemiau esančioje lygtyje:

$$\log(1 + tf_{t,d}) \times idf_t = log_{10} \frac{N}{df_t}$$

čia *N* nurodo dokumentų numerį rinkinyje, $tf_{t,d}$ kaip dažnai žodis *t* pasitaiko dokumente *d*, o idf_t apibūdina paieškos termino svarbą visų kolekcijos dokumentų atžvilgiu.

4 Temų modeliavimas

Temų modeliavimas − tai natūralios kalbos apdorojimo uždavinys, kai duomenų taškai grupuojami į klasterį atsižvelgiant į jų panašumą, o tie, kurie neturi panašumų, bus sugrupuoti į kitus klasterius. Temų modeliavimas apibrėžiamas kaip būdas sugrupuoti duomenis į klasterius taip, kad tame pačiame klasteryje esantys duomenys turėtų daugiau panašumų, palyginti su skirtinguose klasteriuose esančiais duomenimis [8].

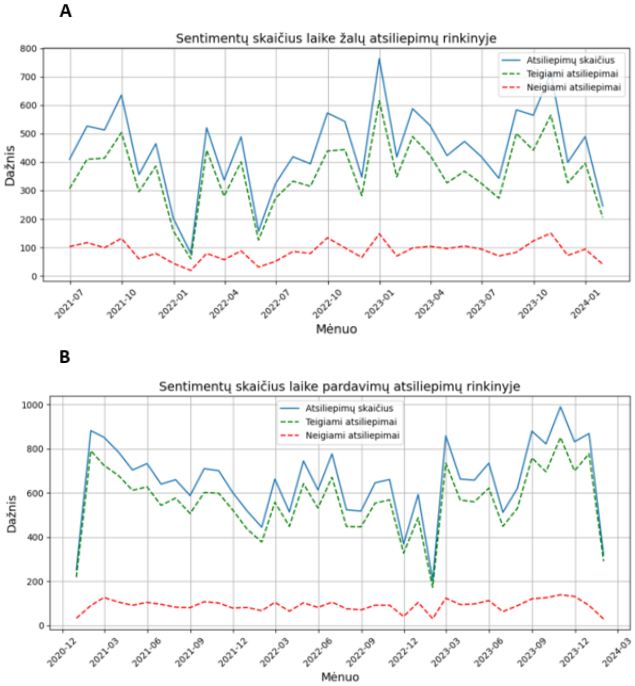
LDA (angl. Latent Dirichlet allocation) yra generuojantis tikimybinis modelis, kurį 2003 m. pirmą kartą pristatė H. Jelodar ir kt [5]. Pagrindinė idėja yra ta, kad dokumentai pateikiami kaip latentinių temų atsitiktiniai mišiniai, kur temą apibūdina žodžių pasiskirstymas. LDA pateikia temas pagal žodžių tikimybes.

5 Eksperimento rezultatai

Norint įvertinti skirtingus vektorizavimo ir klasifikavimo metodus, buvo pasirinkta analizuoti du duomenų rinkinius – atsiliepimai įsigyjant produktą ir klientų vertinimai po draudiminio įvykio atlyginimo. Atlikus pirminį teksto apdorojimą pardavimų atsiliepimų rinkinyje iš viso liko 24662 įrašai, o antrajame duomenų rinkinyje – 14248 atsiliepimai, kurie buvo pateikti klientų po žalos atlyginimo. Abu duomenų rinkiniai turi stulpelį su įvertinimais, pagal kuriuos atsiliepimai buvo suskirstyti į dvi grupes: teigiamus ir neigiamus. Gauta, jog pirmajame duomenų rinkinyje yra 21219 teigiami ir 3443 neigiami atsiliepimai, o antrame rinkinyje – 11478 teigiami ir 2770 neigiami atsiliepimai.

1 paveiksle parodytas bendras, teigiamų ir neigiamų atsiliepimų, paskelbtų kiekvieną mėnesį skirtinguose duomenų rinkiniuose, skaičius. Iš A grafiko, kuris vaizduoja atsiliepimus po žalų atlyginimo matosi, jog didėjant atsiliepimų skaičiui didėja ir neigiami atsiliepimai, tačiau galime pastebėti, jog nuo 2023 metų vasario mėnesio, kai parašomų vertinimų skaičius krito,

o neigiamų atsiliepimų kiekis laikėsi tolygiai. Vertinant abu grafikus pastebima, jog daugiausia klientai vertinimus palieka spalio mėnesiais. Didžiausi kiekiai atsiliepimų A grafiko atveju 2023 sausio mėnesį, o B atveju 2023 metų lapkričio mėn.



1 pav. Žalų atlyginimo (A) ir pardavimų (B) bendrų, neigiamų ir teigiamų atsiliepimų skaičius kas mėnesį.

Neigiamų atsiliepimų temos buvo modeliuojamos kiekvienam ketvirčiui. Norėdami įvertinti sugeneruotų temų kokybę buvo taikomas nuoseklumo (angl. Coherence) balas, kuris padeda nustatyti ar temos yra aiškios ir gerai apibrėžtos. Šis balas remiasi sąsaja tarp žodžių temoje. Pirmiausia, išrenkami svarbiausiai žodžiai kiekvienai temai, o tada parenkamos visos galimos žodžių poros iš atrinktų žodžių. Toliau skaičiuojamas panašumas kiekvienai žodžių porai, matuojant, kaip dažnai du žodžiai pasirodo kartu. Visi panašumai susumuojami kiekvienai temai, gaunant bendrą temos vertę. Galiausiai apskaičiuojamas nuoseklumo balas, dažniausiai kaip vidurkis arba media-

na iš visų temos verčių. Kuo aukštesnis balas, tuo labiau temos apibrėžtos. Gauta, jog didžiausią nuoseklumo balą turinčių temų skaičius yra 10 pardavimų atsiliepimams. 1 lentelėje pateiktos sumodeliuotos temos su raktiniais žodžiais 2023 metų IV ketvirčiui, pardavimų rinkinio atsiliepimams. Iš gautų temų, galima matyti, jog klientai labiausiai nepasitenkinę kainomis, informacijos trūkumu, tikisi geresnio pasiūlymo.

1 lentelė. Pardavimų atsiliepimo rinkinio 2023 metų IV ketvirčiui temų raktiniai žodžiai.

Tema	Raktiniai žodžiai
1	Mažas, draudimas, kainuoti, kaina
2	Pinigai, galėti, kainas, pigiai
3	Galėti, nuolaida, pasiūlymas, draudimas
4	Sunkus, klausimas, įvykis, draudimas
5	Darbuotojas, produktas, turtas, draudimas
6	Paslauga, kaina, šalis, draudimas
7	Darbuotojas, informacija, nežinoti, draudimas
8	Geresnis, bendravimas, kaina, klientas
9	Balas, didelis, draudimas, kaina
10	Bendravimas, žala, trūkti, draudimas

Analogiškas temų modeliavimas atliktas ir antram duomenų rinkiniui, kur didžiausią nuoseklumo balą turinčių temų skaičius yra 3 neigiamų atsiliepimų atveju. Iš 2 lentelėje esančių rezultatų matome, kad daugiausiai tarp temų pasikartoja žodis *draudimas*, *žala*, galime daryti išvadas, kad klientai labiausiai nepatenkinti žalos išmokėjimu, taipogi susidūrė su problemomis naudojantis pakaitiniu automobiliu.

2 lentelė. Atsiliepimų po žalų atlyginimo sumodeliuotų 2023 metų IV ketvirčiui temų raktiniai žodžiai.

Tema	Raktiniai žodžiai
1	Įvykis, draudimas, pakaitinis, automobilis
2	Klientas, žala, kaina, draudimas
3	Žala, darbuotojas, trūkti, draudimas

Sekantis žingsnis - klasifikavimo etapas, naudojant dvi skirtingas vektorizavimo technikas (žodžių maišo ir TF-IDF) bei taikant atraminių vektorių ir naiviojo Bajeso klasifikavimo metodus. Šie klasifikavimo metodus pasirinkti remdamiesi mokslinės literatūros analize ir atliktais tyrimais. Geriausiems rezultatams pasiekti buvo naudojama parametrų gardelė. Kiekvieno modelio rezultatai pateikti 3 ir 4 lentelėse, iš kurių matyti, jog naudojant TF-IDF kartu su naiviuoju Bajeso klasifikatoriumi pasiekiami didžiausi tikslumai abiem duomenų rinkiniams. Mažiausi įverčiai gauti naudojant žodžių atraminių vektorių klasifikatorių.

3 lentelė. Pardavimų atsiliepimų rinkinio klasifikavimo tikslumo metrikos.

Klasifikavimo metodas	Vektorizavimo metodas	Sentimentas	Preciziškumas (%)	Atkūrimas (%)	F1 metrika (%)	Bendras tikslumas
SVM	TF-IDF	Neigiamas	69	45	54	89%
		Teigiamas	91	97	94	
	BOW	Neigiamas	68	44	53	89%
		Teigiamas	91	97	94	
NB	TF-IDF	Neigiamas	73	55	63	91%
		Teigiamas	93	97	95	
	BOW	Neigiamas	66	55	60	90%
		Teigiamas	93	95	94	

4 lentelė. Atsiliepimų rinkinio po žalų atlyginimo klasifikavimo tikslumo metrikos.

Klasifikavimo metodas	Vektorizavimo metodas	Sentimentas	Preciziškumas (%)	Atkūrimas (%)	F1 metrika (%)	Bendras tikslumas
SVM	TF-IDF	Neigiamas	78	63	70	89%
		Teigiamas	91	96	93	
	BOW	Neigiamas	79	60	68	89%
		Teigiamas	90	96	93	
NB	TF-IDF	Neigiamas	82	72	77	91%
		Teigiamas	93	96	95	
	BOW	Neigiamas	76	64	69	89%
		Teigiamas	91	95	93	

6 Išvados

Šiame tyrime buvo naudojami du skirtingi metodai vektorizavimui ir klasifikavimui, siekiant nustatyti klientų atsiliepimų nuotaikas. Taikytos TF-IDF ir žodžių maišo vektorizavimo technikos, kartu su naiviuoju Bajeso ir atraminių vektorių algoritmais. Geriausi rezultatai gauti naudojant naiviojo Bajeso klasifikatorių kartu su TF-IDF, kuris pasiekė net 91% tikslumą skirtingiems duomenų rinkiniams. Atraminių vektorių klasifikatorius pasiekė 89% tikslumą pirmam ir antram duomenų rinkiniams. Lyginant preciziškumo ir atkūrimo metrikas gauta, jog geriausiai kiekvieną klasę atskiria naudojant naiviojo Bajeso kartu su TF-IDF vektorizavimo metodu. Atsiliepimams po produkto įsigijimo geriausiai atskiriama teigiama klasė, tada preciziškumas lygus 93%, o atkūrimo metrika 97%. Atsiliepimams po žalų atlyginimo teigiami sentimentai taip pat geriau atskiriami, gaunamas preciziškumas 93% ir atkūrimo metrika 96%. Nors buvo išbandyti tik keli algoritmai, tolimesniems tyrimams būtų tikslinga išbandyti kitus algoritmus arba kurti hibridinius metodus, siekiant padidinti rezultatų tikslumą. Taip pat atliekant tyrimą pastebėta, kad duomenys nėra balansuoti, todėl ateities darbas bus subalansuoti duomenis. Klientų atsiliepimų nuotaikų nustatymas gali būti naudingas įvairiose srityse. Galimos išmaniosios sistemos, kurios galėtų pateikti vartotojams išsamias produktų, paslaugų ir kt. apžvalgas, nereikalaujant, kad vartotojai peržiūrėtų atskiras apžvalgas, o galėtų tiesiogiai priimtų sprendimus remdamiesi sistemos pateiktais rezultatais.

Literatūra

[1] W. N. I. Al-Obaydy, H. A. Hashim, Y. Najm, and A. A. Jalal. Document classification using term frequency-inverse document frequency and k-means clustering. Indonesian Journal of Electrical Engineering and Computer Science, 27(3):1517–1524, 2022.

[2] A. S. AlQahtani. Product sentiment analysis for amazon reviews. International Journal of Computer Science & Information Technology (IJCSIT) Vol, 13, 2021.

[3] C. Chauhan and S. Sehgal. Sentiment analysis on product reviews. In 2017 International Conference on Computing, Communication and Automation (ICCCA), pages 26–31. IEEE, 2017.

[4] A. Giachanou and F. Crestani. Tracking sentiment by time series analysis. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, pages 1037–1040, 2016.

[5] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. Multimedia tools and applications, 78:15169–15211, 2019.

[6] K. S. Kumar, J. Desai, and J. Majumdar. Opinion mining and sentiment analysis on online customer review. In 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pages 1–4. IEEE, 2016.
[7] G. Murthy, S. R. Allu, B. Andhavarapu, M. Bagadi, and M. Belusonti. Text based sentiment analysis using lstm. Int. J. Eng. Res. Tech. Res, 9(05), 2020.
[8] E. S. Negara, D. Triadi, and R. Andryani. Topic modelling twitter data with latent dirichlet allocation method. In 2019 International Conference on Electrical Engineering and Computer Science (ICECOS), pages 386–390. IEEE, 2019.
[9] A. W. Sari, T. I. Hermanto, and M. Defriani. Sentiment analysis of tourist reviews using k-nearest neighbors algorithm and support vector machine. Sinkron: jurnal dan penelitian teknik informatika, 8(3):1366–1378, 2023.
[10] T. Shivaprasad and J. Shetty. Sentiment analysis of product reviews: A review. In 2017 International conference on inventive communication and computational technologies (ICICCT), pages 298–301. IEEE, 2017.
[11] A. C. M. V. Srinivas, C. Satyanarayana, C. Divakar, and K. P. Sirisha. Sentiment analysis using neural network and lstm. In IOP conference series: materials science and engineering, volume 1074, page 012007. IOP Publishing, 2021.
[12] A. Tripathy, A. Agrawal, and S. K. Rath. Classification of sentimental reviews using machine learning techniques. Procedia Computer Science, 57:821–829, 2015.
[13] Rahat, A. M., Kahir, A., & Masum, A. K. M. (2019, November). Comparison of Naive Bayes and SVM Algorithm based on sentiment analysis using review dataset. In 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART) (pp. 266-270). IEEE.
[14] Limbong, J. J. A. (2022). Analisis Klasifikasi Sentimen Ulasan Pada E-Commerce Shopee Berbasis Word Cloud Dengan Metode Naive Bayes Dan K-Nearest Neighbor (Doctoral dissertation).
[15] Y. Zhang, R. Jin, and Z.-H. Zhou. Understanding bag-of-words model: a statistical framework. International journal of machine learning and cybernetics, 1:43–52, 2010.
[16] Madjid, M. F., Ratnawati, D. E., & Rahayudi, B. (2023). Sentiment Analysis on App Reviews Using Support Vector Machine and Naïve Bayes Classification. Sinkron: jurnal dan penelitian teknik informatika, 8(1), 556-562.