Teksto požymių analizė ir jų efektyvumo vertinimas emocijų klasifikavimo uždavinyje

Ignas Černiauskas, Gražina Korvel

Vilniaus universitetas, Matematikos ir informatikos fakultetas,
Duomenų mokslo ir skaitmeninių technologijų institutas,
Akademijos g. 4, LT-08412 Vilnius, Lietuva
ignas.cerniauskas@mif.vu.lt

Santrauka. Šiame darbe sprendžiamas emocijų atpažinimo iš tekstinių duomenų uždavinys, taikant skirtingus teksto reprezentavimo metodus ir mašininio mokymosi algoritmus. Tyrime naudoti lingvistiniai požymiai, Bag-of-Words ir TF-IDF metodai bei jų kombinacijos. Klasifikavimui taikyti logistinės regresijos, atraminių vektorių mašinos, Naiviojo Bajeso, atsitiktinių miškų ir XGBoost algoritmai. Geriausi rezultatai gauti derinant Bag-of-Words reprezentaciją su papildomais lingvistiniais požymiais ir naudojant logistinės regresijos modelį (tikslumas ir $f_{1\ macro} \approx 0{,}81$). Nustatyta, kad modeliai geriausiai atpažįsta neutralią ir džiaugsmo emocijas, o prasčiausiai – liūdesio.

Raktiniai žodžiai: emocijų klasifikavimas; teksto vektorizavimas; mašininio mokymosi algoritmai.

1 Įvadas

Emocijų atpažinimas iš tekstinių duomenų yra viena iš natūralios kalbos apdorojimo uždavinių, kurio tikslas – identifikuoti ir klasifikuoti žmogaus išreikštas emocijas, tokias kaip džiaugsmas, pyktis ar liūdesys. Emocijų klasifikavimas siejamas su sentimentų analize, tačiau šie du uždaviniai skiriasi savo detalumo lygiu. Sentimentų analizė paprastai nustato tik bendrą teksto toną (teigiamą, neigiamą ar neutralų), o emocijų analizė siekia aptikti detalesnį emocinį turinį, pavyzdžiui, atskirti skirtingas neigiamas emocijas, tokias kaip pyktis ir baimė [2]. Vis dėlto emocijų aptikimas tekste yra sudėtingas uždavinys dėl emocijų raiškos kalboje sudėtingumo, netiesioginių ar metaforinių išraiškų, konteksto priklausomybės bei dėl to, kad tekstiniai duomenys dažnai būna trumpi, neformalūs ir turintys daug žargono [2,5].

Teksto klasifikavimo uždaviniuose pradinis teksto apdorojimas laikomas vienu svarbiausių etapų, turinčių reikšmingą įtaką galutiniam modelių tiks-

DOI: https://doi.org/10.15388/LMITT.2026.6

lumui. Literatūroje pabrėžiama, kad tekstiniai duomenys dažnai pasižymi dideliu triukšmo lygiu, skirtingomis žodžių formomis ir nereikšmingais elementais, todėl jų paruošimas analizei yra būtinas siekiant efektyviau išskirti svarbius lingvistinius ir semantinius dėsningumus [8,3]. Pradinio apdorojimo metu dažniausiai tekstas skaidomas į atskirus vienetus bei taikomi tokie metodai kaip teksto normalizavimas, nereikšmingų žodžių pašalinimas, pagrindinės formos nustatymas (angl. *lemmatization*) ar kamieno išskyrimas (angl. *stemming*) [3]. Tyrimai rodo, kad tinkamai atliktas pradinis teksto apdorojimas gali reikšmingai pagerinti požymių išgavimą iš teksto ir padidinti klasifikavimo modelių tikslumą bei stabilumą [8].

Kitas labai svarbus etapas yra teksto požymių parinkimas, kuris tekstinius duomenis paverčia skaitine forma, tinkama mašininio mokymosi algoritmams. Šiame etape tekstiniai duomenys transformuojami į skaitinius vektorius, kurie atspindi žodžių ar dokumentų savybes ir leidžia modeliams atlikti klasifikavimą. Literatūroje dažniausiai taikomi skirtingi vektorizavimo metodai: Bag of Words, TF-IDF, Word2Vec ir Doc2Vec. Tyrimai rodo, kad tradiciniai metodai, tokie kaip TF-IDF, dažnai pasižymi labai geru našumu teksto klasifikavimo uždaviniuose. Pavyzdžiui, darbe, kuriame buvo sprendžiamas studentų ir dėstytojų komentarų sentimentų klasifikavimo uždavinys [7], nustatyta, kad taikant TF–IDF pagrindu sudarytą teksto požymių reprezentaciją buvo pasiekti geresni klasifikavimo rezultatai nei taikant dvejetainę požymių reprezentaciją,, kai fiksuojamas tik žodžio buvimas arba nebuvimas. Didžiausias tikslumas (iki 97 % trijų klasių klasifikacijoje) pasiektas taikant atsitiktinių miškų algoritmą. Kiti tyrimai taip pat rodo, kad TF-IDF metodas kai kuriais atvejais gali būti efektyvesnis už Word2Vec ir Doc2Vec metodus, ypač sentimentų klasifikavimo uždaviniuose [1].

Mašininio mokymosi metodai plačiai taikomi teksto klasifikavimo uždaviniuose, nes leidžia automatiškai priskirti tekstinius duomenis iš anksto apibrėžtoms kategorijoms. Tyrimuose dažniausiai taikomi tokie klasifikavimo algoritmai kaip atraminių vektorių mašinos (angl. *support vector machines, SVM*), logistinės regresijos modeliai (angl. *logistic regression, LR*), k artimiausių kaimynų metodas (angl. *k-nearest neighbors, k-NN*), Naivusis Bajesas (angl. *Naive Bayes, NB*) bei atsitiktinių miškų algoritmas (angl. *random forest, RF*) [4,6]. Lyginamieji tyrimai rodo, kad skirtingų algoritmų efektyvumas gali skirtis priklausomai nuo naudojamo duomenų rinkinio ir taikomo teksto apdorojimo metodo. Pavyzdžiui, tyrime [4] nustatyta, kad logistinės regresijos ir SVM modeliai pasiekia geriausius rezultatus filmų apžvalgų

duomenų rinkinyje, o k-NN metodas pasižymi didžiausiu tikslumu nepageidaujamų laiškų (angl. *spam*) klasifikavimo uždavinyje. Kiti tyrimai [6] pabrėžia tinkamo išankstinio teksto apdorojimo ir hiperparametrų optimizavimo svarbą, kurie gali reikšmingai pagerinti klasifikavimo modelių veikimą. Taip pat nustatyta, kad pažangesni ansambliniai metodai, tokie kaip gradientinio stiprinimo sprendimų medžiai (angl. *Extreme Gradient Boosting, XGBoost*) ir atsitiktinių miškų algoritmas, pasižymi aukštu klasifikavimo tikslumu tekstinių duomenų analizėje.

Straipsnyje nuosekliai pristatoma tyrimo eiga. Antrame skyriuje aptariama taikyta tyrimo metodologija. Toliau aprašomi naudoti duomenys bei pateikiama leksinių ir morfologinių požymių analizė. Vėlesniame skyriuje nagrinėjami klasifikavimo algoritmų testavimo rezultatai, o baigiamojoje straipsnio dalyje pateikiamos tyrimo išvados.

2 Metodologija

Pradinė duomenų analizė atlikta taikant automatizuotus lietuvių kalbos apdorojimo metodus. Analizei naudotas lietuvių kalbos modelis *spaCy*, leidžiantis atlikti sakinių segmentavimą, pagrindinės formos nustatymą ir kalbos dalių žymėjimą. Tyrimo metu analizuotas skyrybos ženklų pasiskirstymas skirtingose emocinėse kategorijose, vertinta kalbos dalių proporcija tekstuose bei nereikšmingų žodžių dalis. Taip pat apskaičiuota leksinė įvairovė, leidžianti įvertinti žodyno turtingumą ir stilistinius skirtumus tarp emocinių klasių.

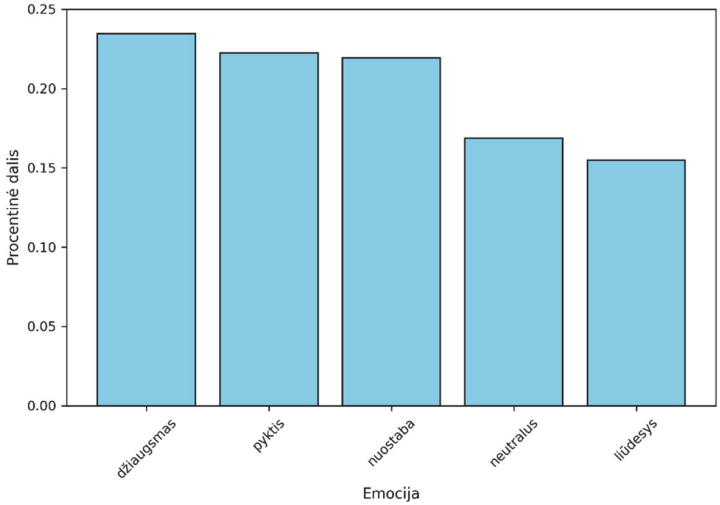
Klasifikavimo etape tekstas buvo skaidomas į pastraipas, išlaikant pirminę dokumento struktūrą. Toks sprendimas pasirinktas siekiant išsaugoti semantinį kontekstą, kadangi pavieniai sakiniai gali būti nepakankamai informatyvūs ar dviprasmiški. Pastraipa laikyta minimaliu kontekstiniu vienetu, kuriame išlaikoma nuosekli minties raida. Tuo tarpu skirtingos pastraipos buvo traktuojamos kaip tarpusavyje nepriklausomi vienetai. Tolimesniame etape, remiantis atlikta lingvistine analize, buvo suformuoti lingvistiniai ir morfologiniai požymiai, tokie kaip nereikšmingų žodžių kiekis, didžiųjų raidžių vartojimas, vidutinis žodžio ilgis, skyrybos ženklų (pvz., šauktukų, klaustukų, kablelių) skaičius bei skirtingų kalbos dalių (pvz., daiktavardžių, veiksmažodžių, būdvardžių) dažniai. Papildomai taikyti du klasikiniai teksto vektorizavimo metodai: „žodžių maišo“ metodas (angl. *Bag of Words - BoW*), grindžiamas žodžių dažnių skaičiavimu, ir TF–IDF metodas

(angl. *Term Frequency – Inverse Document Frequency*). Taip gautos skirtingos teksto reprezentacijos formos, kurios naudotos kaip įvestis klasifikavimo modeliams.

Klasifikavimui taikyti penki algoritmai: logistinė regresija (LR), atraminių vektorių mašina (SVM), naivusis Bajeso algoritmas (NB), atsitiktinių miškų algoritmas (RF) ir XGBoost algoritmas (XGB). Kiekvienam algoritmui hiperparametrai buvo parenkami taikant sisteminę paiešką naudojant *GridSearchCV* metodą kartu su *Pipeline* struktūra. Hiperparametrų paieška vykdyta apibrėžtose reikšmių aibėse, parinktose atsižvelgiant į kiekvieno modelio specifiką: logistinei regresijai optimizuotas reguliacijos stiprumas *C*, reguliacijos tipas (L1 ir L2); SVM modeliui – baudos parametras , nuostolio funkcija ir reguliacijos tipas; naiviajam Bajeso algoritmui – lyginimo parametras *α*, kuris padeda išvengti nulinių tikimybių; XGBoost modeliui – medžių skaičius, maksimalus gylis ir mokymosi greitis, lemiantys modelio sudėtingumą ir mokymosi eigą; o atsitiktinių miškų algoritmui – medžių skaičius, maksimalus gylis bei skaidymo ir lapų formavimo kriterijai (*min_samples_split*, *min_samples_leaf*), turintys įtakos modelio generalizacijai. Hiperparametrų optimizavimas buvo atliekamas taikant 5 kartų kryžminę validaciją, o modelių veikimas vertintas pagal makro vidurkio f_1 rodiklį ($f_{1\ macro}$), kuris yra tinkamas daugiaklasėms, galimai nesubalansuotoms duomenų aibėms, nes visoms klasėms suteikia vienodą svarbą. Duomenys suskaidyti stratifikuotai į mokymo ir testine aibes santykiu 80:20, kur mokymo aibė naudota hiperparametrų optimizavimui, o galutinis modelių vertinimas atliktas su nepriklausoma testine aibe. Eksperimentų metu skirtingos teksto reprezentacijos buvo derinamos su skirtingais klasifikavimo algoritmais, siekiant nustatyti efektyviausią požymių ir modelio kombinaciją.

3 Duomenys ir pradinė analizė

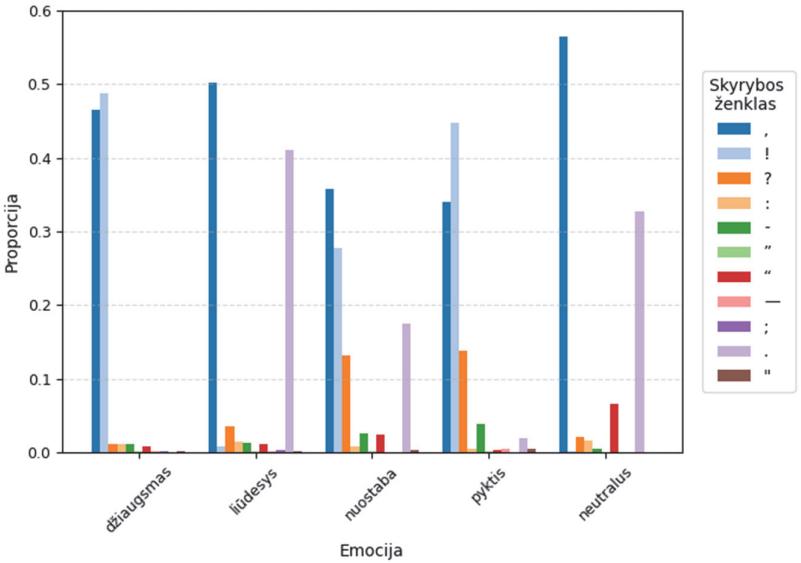
Eksperimentui naudojami Vilniaus universiteto Duomenų mokslo ir skaitmeninių technologijų instituto tyrėjų surinkti duomenys. Bendrą duomenų imtį sudaro 6941 stebėjimas. Duomenyse dominuojančios emocijos yra džiaugsmas, pyktis ir nuostaba (atitinkamai apie 23,5 %, 22 % ir 22 %), tuo tarpu neutrali emocija (17 %) ir liūdesys (15,5 %) sudaro mažesnę dalį (žr. 1 pav.). Nors emocijų pasiskirstymas nėra visiškai tolygus, toks netolygumas laikytinas nedideliu.



1 pav. Duomenų rinkinio emocijų procentinis pasiskirstymas.

Morfologinė analizė rodo, kad skirtingų emocijų tekstuose kalbos dalių pasiskirstymas yra panašus, tačiau galima pastebėti ir tam tikrų dėsningumų. Visose emocijose dažniausiai vartojami daiktavardžiai ir veiksmažodžiai – jie sudaro pagrindą turiniui perteikti ir veiksmui išreikšti. Išsiskiria pykčio tekstai, kuriuose santykinai daugiau įvardžių. Vienas aiškiausių skirtumų šiame duomenų rinkinyje išryškėja analizuojant skyrybos ženklų vartojimą (žr. 2 pav.). Liūdesio ir nuostabos tekstai šiuo požiūriu gana panašūs – jų skyryba tolygesnė, be ryškaus emocinių ženklų dominavimo. Džiaugsmo ir pykčio emocijose, priešingai, pasižymi labai panašia raiška: jose gausu šauktukų (!), o taškai (.) vartojami retai. Tai sukuria stipresnės emocinės įtampos ir spontaniškumo įspūdį. Taip pat pastebėtina, kad nuostabos ir pykčio tekstuose dažniau pasitaiko klaustukų (?), palyginti su kitomis emocijomis, kas sietina su nustebimo, abejonės ar retorinio klausimo išraiška.

Nagrinėjant nereikšmingus žodžius (žr. 1 lentelė.), nustatyta, kad didžiausia vidutinė nereikšmingų žodžių dalis yra liūdesio (0,247 ± 0,155) ir pykčio (0,241 ± 0,195) tekstuose, o mažiausia – nuostabos (0,207 ± 0,176) ir neutralios emocijos (0,214 ± 0,130) tekstuose. Vis dėlto skirtumai tarp emocijų nėra dideli, o aukšti standartinio nuokrypio dydžiai rodo didelę sakinių įvairovę kiekvienoje emocijoje.



2 pav. Skyrybos ženklų vartojimas skirtingose emocijų tekstuose.

1 lentelė. Nereikšmingų žodžių santykinė dalis ir leksinė įvairovė pagal emocijas.

Emocija	Nereikšmingų žodžių santykis (vid ± sd)	Leksinė įvairovė
Džiaugsmas	0,221 ± 0,158	0,450
Liūdesys	0,247 ± 0,155	0,529
Nuostaba	0,207 ± 0,176	0,521
Pyktis	0,241 ± 0,195	0,531
Neutralus	0,214 ± 0,130	0,431

Tai pat, nagrinėtas leksinės įvairovės rodiklis parodo, kiek skirtingų leksinių vienetų vartojama kiekvienos emocijos tekstuose (žr. 1 lentelė.), todėl didesnė reikšmė reiškia įvairesnį žodyną, o mažesnė – didesnį pasikartojimą. Gauti rezultatai rodo, kad pykčio ir liūdesio tekstai pasižymi didžiausia leksine įvairove, o džiaugsmo ir ypač neutralūs tekstai yra labiau pasikartojantys.

4 Klasifikavimo rezultatai

Klasifikavimo eksperimentų rezultatai rodo, kad skirtingi tekstų reprezentavimo metodai ir lingvistiniai požymiai turi įtakos modelių veikimui (žr. 2 lentelė.). Vertinant modelius, naudotus tik su lingvistiniais požymiais, geriausi rezultatai buvo gauti taikant XGBoost modelį (tikslumas apie 0,65, $f_{1\ macro}$ apie 0,65). Tuo tarpu prasčiausi rezultatai šioje grupėje buvo gauti taikant Naivio-jo Bajeso modelį, kurio klasifikavimo kokybė buvo žemiausia tarp nagrinėtų algoritmų.

2 lentelė. Teksto klasifikavimo modelių rezultatų palyginimas pagal testavimo tikslumą ir $f_{1\ macro}$.

Teksto reprezentacija	Klasifikavimo algoritmas	Testavimo aibės tikslumas	Testavimo aibės $f_{1\ macro}$
Lingvistiniai požymiai	LR	0,6091	0,6004
	SVM	0,6055	0,5937
	NB	0,5536	0,5497
	XGB	0,6515	0,6494
	RF	**0,6156**	**0,6063**
TF-IDF	LR	0,7423	0,7386
	SVM	0,7502	0,7456
	NB	0,7394	0,7332
	XGB	0,6393	0,6352
	RF	0,4240	0,3502
BoW	LR	0,7279	0,7265
	SVM	0,7264	0,7249
	NB	0,7365	0,7305
	XGB	0,6530	0,6504
	RF	0,4140	0,3438
TF-IDF + Lingvistiniai požymiai	LR	0,7963	0,7967
	SVM	0,7869	0,7881
	NB	0,7156	0,7133
	XGB	0,7516	0,7497
	RF	0,5313	0,4577
BoW + Lingvistiniai požymiai	LR	**0,8056**	**0,8058**
	SVM	**0,7977**	**0,7990**
	NB	**0,7768**	**0,7797**
	XGB	**0,7559**	**0,7562**
	RF	0,5292	0,4562

Naudojant klasikinius teksto vektorizavimo metodus, tokius kaip Bag-of-Words ir TF-IDF, modelių veikimo kokybė pastebimai pagerėjo. Geriausi rezultatai šioje kategorijoje buvo gauti taikant SVM modelį su TF-IDF reprezentacija (tikslumas apie 0,75, $f_{1\ macro}$ apie 0,75). Tuo tarpu prasčiausi rezultatai buvo gauti taikant atsitiktinių miškų modelį, kuriam tiek BoW, tiek TF-IDF reprezentacija buvo neefektyvi, o klasifikavimo tikslumas nesiekė 0,43.

Geriausi rezultatai buvo gauti derinant teksto vektorizavimo metodus su papildomais lingvistiniais požymiais. Aukščiausia klasifikavimo kokybė nustatyta taikant logistinės regresijos modelį su Bag-of-Words ir papildomais požymiais, kai tikslumo ir $f_{1\ macro}$ rodikliai pasiekė 0,81. Tuo tarpu prasčiausi rezultatai šioje modelių grupėje buvo gauti taikant atsitiktinių miškų modelį su TF-IDF ir papildomais požymiais, kurio tikslumas išliko ženkliai žemesnis nei kitų algoritmų.

Analizuojant detaliau (žr. 3 lentelė.), kurioje pateikti geriausiai veikusio modelio klasifikavimo rodikliai pagal emocijas, matyti, kad geriausi rezultatai gaunami neutralios emocijos tekstams, nes šiai emocijai f_1 reikšmė yra didžiausia (0,8590). Taip pat aukšti įverčiai džiaugsmo emocijai (f_1 = 0,8546). Tuo tarpu prasčiausi įverčiai nustatyti pykčio emocijai (f_1 = 0,7573). Liūdesio ir nuostabos emocijoms būdingi vidutiniai rezultatai – jų f_1 įverčiai (atitinkamai 0,7832 ir 0,7752) patenka tarp aukščiausių ir žemiausių reikšmių. Tokius rezultatus galima paaiškinti tuo, kad neutralūs tekstai pasižymi mažesniu lingvistiniu kintamumu ir paprastesne struktūra, todėl modeliams juos lengviau atpažinti. Džiaugsmo emocija taip pat dažnai išreiškiama aiškiais požymiais, pavyzdžiui, būdingais žodžiais ar skyrybos ženklais (pvz., šauktukais). Tuo tarpu pykčio, liūdesio ir nuostabos emocijos pasižymi didesne leksine įvairove bei subtilesne raiška, todėl jų klasifikavimas yra sudėtingesnis.

3 lentelė. Geriausiai veikusio modelio klasifikavimo rodikliai pagal emocijas.

Emocija	**Tikslumas (angl. *Precision*)**	**Atkūrimas (angl. *Recall*)**	f_1
Džiaugsmas	0,8358	0,8742	0,8546
Liūdesys	0,7850	0,7814	0,7832
Nuostaba	0,7938	0,7574	0,7752
Pyktis	0,7573	0,7573	0,7573
Neutralus	0,8590	0,8590	0,8590

Apibendrinant galima teigti, kad vien tik lingvistinių požymių naudojimas lemia ribotus klasifikavimo rezultatus, tačiau žymiai geresni rezultatai gaunami taikant klasikinius teksto vektorizavimo metodus. Vis dėlto aukščiausias klasifikavimo tikslumas nustatytas derinant Bag-of-Words arba TF-IDF reprezentaciją su papildomais lingvistiniais požymiais, kas rodo, kad skirtingų informacijos šaltinių integravimas leidžia efektyviau atskirti emocines tekstų kategorijas.

5 Išvados

Šiame darbe buvo sprendžiamas emocijų atpažinimo iš tekstinių duomenų uždavinys., kuris laikomas sudėtingesne sentimentų analizės forma, nes siekiama identifikuoti konkrečias emocijas, o ne tik bendrą teksto toną.

Morfologinė analizė atskleidė, kad visuose tekstuose dominuoja daiktavardžiai ir veiksmažodžiai, tačiau pykčio tekstuose santykinai dažniau vartojami įvardžiai. Analizuojant skyrybos ženklus nustatyta, kad džiaugsmo ir pykčio tekstuose dažnai naudojami šauktukai, o neutralios ir liūdesio emocijų tekstuose – taškai. Taip pat pastebėta, kad liūdesio ir pykčio tekstuose yra didesnė nereikšmingų žodžių dalis. Leksinės įvairovės analizė parodė, kad pykčio ir liūdesio tekstai pasižymi didžiausia leksine įvairove, o neutralūs ir džiaugsmo tekstai pasižymi mažesne leksine įvairove, todėl juose dažniau kartojasi tie patys žodžiai.

Klasifikavimo eksperimentų rezultatai parodė, kad vien lingvistinių požymių naudojimas lemia ribotus klasifikavimo rezultatus. Naudojant TF-IDF arba Bag-of-Words reprezentaciją, dalies modelių rezultatai pagerėjo, ypač taikant atraminių vektorių ir logistinės regresijos algoritmus. Vis dėlto ne visais atvejais buvo gauti geresni rezultatai – taikant atsitiktinių miškų algoritmą, aukščiausi rezultatai nustatyti naudojant tik lingvistinius požymius. Geriausi bendri rezultatai buvo gauti derinant teksto vektorizavimo metodus su papildomais lingvistiniais požymiais.

Aukščiausią klasifikavimo tikslumą (0,8056) ir $f_{1\ macro}$ (0,8058) rodiklį pasiekė logistinės regresijos modelis su Bag-of-Words ir lingvistiniais požymiais. Detalesnė analizė parodė, kad pagal f_1 rodiklį modelis geriausiai atpažįsta neutralios emocijos (0,8590) tekstus ir tiksliai identifikuoja džiaugsmo emociją, tuo tarpu prasčiausiai atpažįstama emocija yra pyktis (0,7573), o nuostabos ir liūdusio emocijos atpažįstamos vidutiniškai.

Literatūra

[1] Abubakar, H. D., Umar, M., & Bakale, M. A. (2022). Sentiment classification: Review of text vectorization methods: Bag of words, Tf-Idf, Word2vec and Doc2vec. *SLU Journal of Science and Technology*, *4*(1), 27-33.

[2] Acheampong, F. A., Wenyu, C., & Nunoo-Mensah, H. (2020). Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, *2*(7), e12189.

[3] HaCohen-Kerner, Y., Miller, D., & Yigal, Y. (2020). The influence of preprocessing on text classification using a bag-of-words representation. *PloS one*, *15*(5), e0232525.

[4] Hassan, S. U., Ahamed, J., & Ahmad, K. (2022). Analytics of machine learning-based algorithms for text classification. *Sustainable Operations and Computers, 3, 238-248*.

[5] Nandwani, P., & Verma, R. (2021). A review on sentiment analysis and emotion detection from text. *Social network analysis and mining*, *11*(1), 81.

[6] Occhipinti, A., Rogers, L., & Angione, C. (2022). A pipeline and comparative study of 12 machine learning models for text classification. *Expert Systems with Applications*, *201*, 117193.

[7] Rakhmanov, O. (2020). A comparative study on vectorization and classification techniques in sentiment analysis to classify student-lecturer comments. *Procedia Computer Science*, *178*, 194-204.

[8] Siino, M., Tinnirello, I., & La Cascia, M. (2024). Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers. *Information Systems*, *121*, 102342.