



Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Aplinkosaugos, socialinės ir valdymo informacijos sentimentų analizė socialinėje medijoje

Baigiamasis magistro studijų projektas

Goda Sabaliauskienė

Projekto autorė

Doc. dr. Mantas Landauskas

Vadovas

Prof. dr. Lina Dagilienė

Vadovė

Kaunas, 2024



Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Aplinkosaugos, socialinės ir valdymo informacijos sentimentų analizė socialinėje medijoje

Baigiamasis magistro studijų projektas

Didžiųjų verslo duomenų analitika (6213AX001)

Goda Sabaliauskienė

Projekto autorė

Prof. dr. Lina Dagilienė

Vadovė

Doc. dr. Mantas Landauskas

Vadovas

Doc. dr. Viktorija Varaniūtė

Recenzentė

Prof. dr. Tomas Ruzgas

Recenzentas

Kaunas, 2024



Kauno technologijos universitetas

Matematikos ir gamtos mokslų fakultetas

Goda Sabaliauskienė

Aplinkosaugos, socialinės ir valdymo informacijos sentimentų analizė socialinėje medijoje

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 sąrašė;
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.

Goda Sabaliauskienė

Patvirtinta elektroniniu būdu

Sabaliauskienė, Goda. Aplinkosaugos, socialinės ir valdymo informacijos sentimentų analizė socialinėje medijoje. Magistro studijų baigiamasis projektas / vadovas doc. dr. Mantas Landauskas, vadovė prof. dr. Lina Dagilienė; Kauno technologijos universitetas, Matematikos ir gamtos mokslų fakultetas.

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

Reikšminiai žodžiai: ESG, aplinkosauga, socialinė sritis, valdymas, sentimentų analizė, teksto analizė, mašininis mokymasis, VADER, socialinė žiniasklaida.

Kaunas, 2024. 69 p.

Santrauka

ESG, apimantis aplinkosaugos, socialinės atsakomybės ir valdymo aspektus, yra vis svarbesnis įmonių veiklos vertinimo rodiklis. Informacinės technologijos ir interneto plėtra suteikia galimybę greitai ir plačiai skleisti informaciją, o socialinė žiniasklaida tampa pagrindiniu kanalu, per kurį visuomenė gali išreikšti savo nuomonę.

Projektas ne tik atspindi aplinkosaugos, socialinės atsakomybės ir valdymo informacijos atskleidimo svarbą, bet ir išryškina socialinės žiniasklaidos vaidmenį formuojant visuomenės požiūrį į įmonių strategijas. Tyrimas atskleidžia, kad sentimentų analizė yra naudingas įrankis ne tik konkurentų ir vartotojų požiūrio išsiaiškinimui, bet ir ESG praktikų vertinimui. Tačiau šioje srityje dar trūksta mokslinių tyrimų, todėl šis darbas yra svarbus indėlis į šios analizės pritaikymą mados sektoriuje.

Darbe naudojamos įvairios metodologijos, įskaitant literatūros analizę ir kiekybinę duomenų analizę, naudojant NLP (*Natural Language Processing*) ir mašininio mokymosi algoritmus, skirtus teksto sentimentų identifikavimui. Taip pat sukurta vartotojo sąsaja Microsoft Power BI pagrindu, leidžianti vizualizuoti ir interpretuoti sentimentų analizės rezultatus interaktyviose švieslentėse. Analizuojant **E**, **S**, **G** kiekvieną sritį atskirai, sentimentų poliarškumui vertinti atlikta VADER sentimentų analizė, panaudotas Naivaus Bajeso metodas bei atraminių vektorių metodas, kur keičiant **C** parametro reikšmes, surastas tinkamiausias modelis.

Atlikus tyrimą pastebėta, kad *Twitter* skelbiamos žinutėse apie mados sektorių dominuoja teigiamos nuotaikos. Galima teigti, kad visuomenė gana palankiai vertina kompanijų pastangas skirtas aplinkosaugai ir valdymo skaidrumo temoms. Daugiau negatyvių žinučių pasitaikė socialinės atsakomybės tema, todėl siekdamos išlaikyti savo reputaciją, kompanijos turėtų atsižvelgti į šį aspektą.

Sabaliauskienė, Goda. Sentiment Analysis of Environmental, Social and Governance Information on Social Media. Master's Final Degree Project / supervisor Assoc. Prof. Mantas Landauskas, supervisor Prof. Dr. Lina Dagilienė; Faculty of Mathematics and Natural Sciences, Kaunas University of Technology.

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

Keywords: ESG, Sentiment Analysis, Text Analysis, Machine Learning (ML), VADER, Social Media

Kaunas, 2024. 69 pages.

Summary

ESG, which covers environmental, social responsibility and governance aspects, is an increasingly important indicator for measuring corporate performance. Information technology and the development of the internet allow information to be disseminated quickly and widely, and social media is becoming the main channel through which the public can express its views.

The project not only reflects the importance of environmental, social and governance information disclosure, but also highlights the role of social media in shaping public attitudes towards corporate strategies. The study reveals that sentiment analysis is a useful tool not only for understanding competitors and consumers attitudes, but also for assessing ESG practices. However, there is still a lack of research in this area and this paper makes an important contribution to the application of sentiment analysis in the fashion industry.

The work employs a variety of methodologies, including literature review and quantitative data analysis, using NLP (Natural Language Processing) and machine learning algorithms to identify textual sentiments. A user interface via Microsoft Power BI has also been developed to visualise and interpret the results of the sentiment analysis on interactive dashboards. Analysing E, S, G in each domain separately, a VADER sentiment analysis was performed to assess the polarity of sentiments. The Naive Bayes method and the SVM method were used to find the most appropriate model by varying the values of parameter C.

The study has shown that *Twitter* posts about the fashion&apparel sector are dominated by positive sentiment. It can be assumed that the public is quite supportive of companies efforts to deal with environmental and governance transparency issues. More negative messages were observed on the topic of social responsibility, and companies should take this aspect into consideration in order to maintain their reputation.

Turinys

Lentelių sąrašas	7
Paveikslų sąrašas	8
Santrumpų ir terminų sąrašas	10
Įvadas.....	11
1. Literatūros apžvalga	13
1.1. Aplinkosaugos, socialinės ir valdymo informacijos atskleidimas ir analizė.....	13
1.1.1. ESG atsiradimas ir termino samprata	13
1.1.2. Aplinkosaugos, socialinės ir valdymo informacijos aktualumas mados sektoriuje	17
1.1.3. ESG tema socialinėje žiniasklaidoje	20
1.2. Sentimentų analizė.....	23
1.2.1. Sentimentų analizės pagrindiniai aspektai.....	23
1.2.2. Sentimentų analizės tyrimo problematika	25
1.3. Teksto apdorojimas ir sentimentų analizės metodai.....	26
1.4. Alternatyvūs būdai sentimentų analizei atlikti	28
1.5. Literatūros analizės apibendrinimas	30
2. Tyrimų metodologija.....	31
2.1. Duomenų gavyba.....	31
2.2. Teksto apdorojimas	32
2.3. Duomenų balansavimo metodai	34
2.4. Leksikonu grįsti metodai	35
2.5. Taisyklėmis grįsti metodai	36
2.6. Mašininio mokymosi metodai	37
2.6.1. Atraminių vektorių klasifikatorius (SVM)	38
2.6.2. Naivusis Bajesas klasifikatorius (NB).....	41
2.7. Tikslumo vertinimo metrikos	42
3. Tyrimo eiga ir rezultatai.....	44
3.1. Duomenų gavybos procesas	44
3.2. Teksto apdorojimas ir požymių išskyrimas.....	46
3.3. Sentimentų analizė naudojant <i>Twitter</i> žinutes.....	50
3.3.1. VADER sentimentų analizė	50
3.3.2. Žinučių analizė atraminių vektorių klasifikatoriumi (SVM).....	53
3.3.3. Žinučių analizė Naivaus Bajeso (NB) metodu	55
3.4. Rezultatų vertinimas pritaikius duomenų balansavimo metodus	56
3.4.1. VADER sentimentų analizė su subalansuota imtimi.....	56
3.4.2. Žinučių analizė atraminių vektorių metodu su subalansuota imtimi.....	57
3.4.3. Žinučių analizė Naivaus Bajeso metodu su subalansuota imtimi	58
3.5. Tyrimo rezultatų apžvalga.....	58
3.6. Rekomendacijos	60
Išvados	62
Literatūros sąrašas	64
Priedai.....	70
1 priedas. Dažniausiai paminėti ESG terminai [14]	70
2 priedas. Vartotojo sąsaja parengta pagal VADER sentimentų analizės metodą Power BI analitikos įrankiu	71

Lentelių sąrašas

1 lentelė. Sentimentų analizės tyrimai iš socialinės žiniasklaidos duomenų [2, 77, 78].....	22
2 lentelė. Matematinį metodų tikslumas naudojant mašininio mokymo algoritmus [48].....	27
3 lentelė. <i>Twitter</i> duomenų gavybos politika	31
4 lentelė. Teksto apdorojimo etapai su atsitiktinai pasirinktomis dviem <i>Twitter</i> žinutėmis	47
5 lentelė. VADER euristicų panaudojimas su atsitiktinai pasirinkta <i>Twitter</i> žinute.....	50
6 lentelė. Žinučių statistiniai duomenys aplinkosaugos (<i>E</i>) tema.....	51
7 lentelė. Žinučių statistiniai duomenys socialinės atsakomybės (<i>S</i>) tema.....	51
8 lentelė. Žinučių statistiniai duomenys valdymo (<i>G</i>) tema	52
9 lentelė. Žinučių metrikų vertinimas VADER metodu	53
10 lentelė. Žinučių metrikų vertinimas atraminių vektorių metodu aplinkosaugos (<i>E</i>) tema.....	54
11 lentelė. Žinučių metrikų vertinimas atraminių vektorių metodu socialinės atsakomybės (<i>S</i>) tema	54
12 lentelė. Žinučių metrikų vertinimas atraminių vektorių metodu valdymo (<i>G</i>) tema	55
13 lentelė. Žinučių metrikų vertinimas Naivaus Bajeso metodu	56
14 lentelė. Žinučių metrikų vertinimas VADER metodu pritaikius imties didinimo techniką.....	56
15 lentelė. Žinučių metrikų vertinimas atraminių vektorių metodu aplinkosaugos (<i>E</i>) tema pritaikius imties didinimo techniką	57
16 lentelė. Žinučių metrikų vertinimas atraminių vektorių metodu valdymo (<i>G</i>) tema pritaikius imties didinimo techniką	57
17 lentelė. Žinučių metrikų vertinimas Naivaus Bajeso metodu pritaikius imties didinimo techniką	58
18 lentelė. Apibendrintas žinučių metrikų vertinimas VADER metodu.....	59
19 lentelė. Geriausių rezultatų palyginimas gautas atraminių vektorių metodu	59
20 lentelė. Apibendrintas žinučių metrikų vertinimas Naivaus Bajeso metodu	60
21 lentelė. Iššūkiai ir pasiūlymai įvairiems sentimentų analizės etapams	61

Paveikslų sąrašas

1 pav. ESG ataskatų pateikimo dinamika S&P 500 kompanijų [12]	13
2 pav. <i>Google Trends</i> : susidomėjimas ESG tema pasaulyje [13]	14
3 pav. <i>Google Trends</i> : susidomėjimas ESG tema Lietuvoje [13]	14
4 pav. ESG pagrindinės sritys	15
5 pav. ESG temų pasiskirstymas pagal sektorius [14]	16
6 pav. Tiekimo grandinės schema. Parengta pagal [23]	18
7 pav. Sertifikatai pagal sritis. Parengta pagal [39, 40]	20
8 pav. <i>Twitter</i> vartotojų pasiskirstymas 2023 m. duomenimis	21
9 pav. Sentimentų analizės panaudojimas. Parengta pagal [49, 51]	23
10 pav. Sentimentų analizės lygmenys [49, 51, 52]	24
11 pav. Bendrinė sentimentų analizės atlikimo eiga [49]	25
12 pav. Sentimentų analizės procesas [59]	26
13 pav. Sentimentų klasifikavimo metodai. Parengta pagal [49, 57]	27
14 pav. <i>MonkeyLearn</i> sentimentų analizės pavyzdinė švieslentė [52]	28
15 pav. <i>Brand24</i> socialinės žiniasklaidos žinučių sentimentų analizės pavyzdinė švieslentė [54] ..	29
16 pav. Teksto apdorojimo pavyzdys, taikant lemantizavimą ir kamienizavimą	32
17 pav. Atraminų vektorių klasifikatorius: a) 2 klasių klasifikavimo pavyzdys, b) daugiau kaip dviejų klasių klasifikavimo pavyzdys [45]	39
18 pav. Klasifikavimas Naivaus Bajeso metodu [80]	41
19 pav. Sumaišymo matricos pavyzdys 2 klasėms	43
20 pav. Sumaišymo matricos pavyzdys 3 klasėms	43
21 pav. <i>Twitter</i> žinutės pavyzdys aplinkosaugos tema	45
22 pav. <i>Twitter</i> žinutės pavyzdys paruoštas teksto apdorojimui	45
23 pav. <i>Twitter</i> žinučių pavyzdžiai socialinės atsakomybės tema	46
24 pav. <i>Twitter</i> žinučių pavyzdžiai valdymo tema	46
25 pav. Dažniausiai minimų žodžių stulpelinė pasiskirstymo diagrama: a) aplinkosaugos srityje, b) socialinės atsakomybės srityje, c) valdymo srityje	48
26 pav. Žodžių debesis, sudarytas iš dažniausiai paminėtų žodžių: a) aplinkosaugos srityje, b) socialinės atsakomybės srityje, c) valdymo srityje	48
27 pav. Atstumų matrica su raktiniais žodžiais aplinkosaugos (<i>E</i>) tema	49
28 pav. Atstumų matrica su raktiniais žodžiais socialinės atsakomybės (<i>S</i>) tema	49
29 pav. Atstumų matrica su raktiniais žodžiais valdymo (<i>G</i>) tema	50
30 pav. Žinučių pasiskirstymas tarp neigiamų, neutralių ir teigiamų skirtingose srityse: a) aplinkosauga, b) socialinė atsakomybė, c) valdymas	52
31 pav. Duomenų pasiskirstymo kiekis su trendo linija: a) aplinkosauga, b) socialinė atsakomybė, c) valdymas	52
32 pav. 5-erių metų teigiamų ir neigiamų žinučių kiekis: a) aplinkosauga, b) socialinė atsakomybė, c) valdymas	53
33 pav. Žinučių metrikų vertinimas atraminų vektorių metodu su skirtingomis parametro <i>C</i> reikšmėmis aplinkosaugos (<i>E</i>) tema, Oy ašyje – teisingai suklasifikuotų žinučių dalis: a) neigtyvios, b) neutralios, c) teigiamos	54
34 pav. Žinučių metrikų vertinimas atraminų vektorių metodu su skirtingomis parametro <i>C</i> reikšmėmis socialinės atsakomybės (<i>S</i>) tema, Oy ašyje – teisingai suklasifikuotų žinučių dalis: a) neigtyvios, b) neutralios, c) teigiamos	55

35 pav. Žinučių metrikų vertinimas atraminių vektorių metodu su skirtingomis parametro <i>C</i> reikšmėmis valdymo (G) tema, Oy ašyje – teisingai suklasifikuotų žinučių dalis: a) negatyvios, b) neutralios, c) teigiamos.....	55
36 pav. Žinučių metrikų vertinimas atraminių vektorių metodu su skirtingomis parametro <i>C</i> reikšmėmis aplinkosaugos (E) tema, Oy ašyje – teisingai suklasifikuotų žinučių dalis: a) negatyvios, b) neutralios, c) teigiamos	57
37 pav. Žinučių metrikų vertinimas atraminių vektorių metodu su skirtingomis parametro <i>C</i> reikšmėmis valdymo (G) tema, Oy ašyje – teisingai suklasifikuotų žinučių dalis: a) negatyvios, b) neutralios, c) teigiamos.....	58

Santrumpų ir terminų sąrašas

Santrumpos:

ESG – Aplinkosaugos (angl. *Environmental*), socialinės atsakomybės (angl. *Social*) ir valdymo (angl. *Governance*) sričių derinys;

CSRD – įmonių tvarumo ataskaitų teikimo direktyva (angl. *Corporate Sustainability Reporting Directive*);

NLP – natūralios kalbos apdorojimas (angl. *Natural Language Processing*);

TF – terminų dažnis (angl. *Term Frequency*);

IDF – atvirkštinis dokumentų dažnis (angl. *Inverse Document Frequency*);

TF-IDF – terminų dažnio-atvirkštinio dokumentų dažnio metodas (angl. *Term Frequency-Inverse Document Frequency*);

VADER (angl. *Valence Aware Dictionary and sEntiment Reasoner*) – leksikonu ir taisyklėmis grįstas sentimentų analizės metodas;

ML – Mašininis mokymasis (angl. *Machine learning*);

SVM – Atraminių vektorių klasifikatorius (angl. *Support Vector Machine*);

NB – Naivaus bajeso klasifikatorius (angl. *Naïve Bayes*);

DI – dirbtinis intelektas.

Terminai:

Sentimentų analizė (SA) – nuomonių, nuotaikų analizė – yra duomenų gavybos metodas, leidžiantis interpretuoti ir klasifikuoti emocijas (teigiamas, neigiamas ir neutralias) naudojant tekstinius duomenis;

Vartotojo sąsaja – tai Microsoft Power BI pagrindu sukurtos vizualios ir interaktyvios švieslentės;

Power BI – tai Microsoft įrankis, skirtas apjungti, apdoroti duomenis, atlikti skaičiavimus ir pateikti švieslentes su besinaujinančiais duomenimis.

Ivadas

Temos aktualumas. Aplinkosaugos (angl. *Environmental*), socialinės atsakomybės (angl. *Social*) ir valdymo (angl. *Governance*), toliau ESG informacijos atskleidimas įgauna vis didesnę pagreitį verslo pasaulyje. ESG – tai įmonių nefinansinių rodiklių visuma – aplinkosaugos, socialinės atsakomybės ir valdymo srityse [1, 30].

Klimato kaita, didėjanti oro tarša, pandemijos, darbuotojų reikalavimas saugesnių darbo sąlygų ir gerovės, rizikų valdymas, etiniai principai bei skaidrumo siekimas verčia kompanijas imtis permainų savo veiklose. Informacinių technologijų plėtra ir lengvas informacijos pasiekiamumas suteikia galimybes kompanijoms skelbti apie pokyčius, pateikti nefinansinę informaciją interneto svetainėse ir socialinėje žiniasklaidoje. Šių dienų socialinė žiniasklaida yra labai dinamiška ir įtrauki. Informacijos pasiekiamumas yra labai greitas, žmonės noriai komentuoja [2].

ESG informacijos atskleidimas svarbus investicijų pritraukimui bei rizikos valdymui, didina įmonės reputaciją ir skaidrumą, užtikrina reguliavimo atitikties laikymąsi, pritraukia jaunus darbuotojus, kuriems svarbi socialinė ir aplinkosauginė atsakomybė [30]. Tačiau ne mažiau aktualu išsiaiškinti, ar įmonių atstovai ir visuomenė vienodai vertina įmonių pastangas. Visuomenės nuomonei įvertinti plačiai taikomas sentimentų analizės metodas [2, 48]. Sentimentų analizės metodas yra gerai išnagrinėtas ir dažnai naudojamas konkurentų, klientų kuriamam turiniui vertinti, produktų bei paslaugų atsiliepimams įvertinti [44, 48], tačiau ne tiek daug randama straipsnių, kuriuose nagrinėjami sentimentai ESG tema.

Šiame darbe analizuojamos *Twitter* žinutės, susijusios su mados pramonės prekiniais ženklais. Mados pramonė visoje tiekimo grandinėje susiduria su dideliais aplinkosauginiais bei socialiniais iššūkiais [24, 26, 32, 35, 37], todėl visuomenės nuomonė yra labai svarbi. Įmonės, nustatydamos savo tikslus ir publikuodamos ESG informaciją, įsipareigoja darbuotojams, investuotojams ir visuomenei. Kompanijoms svarbu neprarasti ryšio su visais suinteresuotais asmenimis, todėl aktualu ne tik publikuoti informaciją, bet ją analizuoti bei įvertinti per visuomenės nuomonės prizmę.

Svarbu paminėti, kad tiek ESG informacijos atskleidimas, tiek ir vartotojų reakcijos tyrimas pasinaudojus sentimentų analize yra aktuali šių dienų tema. Šių metodų taikymas padeda palaikyti kontaktą su suinteresuotais asmenimis ir skatinti tvaraus verslo vystymąsi.

Darbo problema. Tvarumo ataskaitas teikia vis daugiau įmonių [12]. Taip bandoma patenkinti suinteresuotųjų šalių poreikį apie įmones žinoti daugiau nei tik finansinius duomenis. Šiame tyrime siekiama atskleisti ne pačių įmonių teikiamų ataskaitų turinį, bet visuomenės požiūrį į kompanijų atsakomybės aplinkosaugos, socialinės ir valdymo temomis.

Mados pramonė yra labai jautri visuomenės nuomonei, stipriai priklauso nuo vartotojų įpročių bei elgsenos, todėl kompanijoms labai svarbu suprasti vartotojų lūkesčius, nuotaikas ir turėti prieinamą būdą stebėti šiuos pokyčius.

Objektas – socialinės žiniasklaidos vartotojų (visuomenės) nuomonė ESG tema – žinutės *Twitter* tinkle.

Tyrimo tikslas – ištirti mados pramonės ESG informacijos atskleidimą socialinėje žiniasklaidoje taikant sentimentų analizės metodus.

Tyrimo uždaviniai:

1. apžvelgti aplinkosaugos, socialinės ir valdymo temos aktualumą pasaulyje;
2. pagrįsti mados pramonės įmonių ESG informacijos svarbą socialinėje žiniasklaidoje ir išanalizuoti kokia informacija apie mados pramonės įmones atskleidžiama socialinės žiniasklaidos kanale *Twitter*;
3. išsiaiškinti, kurie matematiniai metodai geriausiai tinkami sentimentų analizei;
4. taikant matematinius metodus nustatyti, ar socialinėje žiniasklaidoje skleidžiama informacija yra pozityvi, negatyvi ar neutrali;
5. sukurti vartotojo sąsają, kurioje sentimentų analizės rezultatai būtų lengvai interpretuojami ir turėtų praktinį panaudojimą verslo įmonėse.

Tyrimo metodai

1. **Literatūros apžvalga.** Analizuojami moksliniai straipsniai, susiję su ESG informacijos atskleidimu bei sentimentų analizės taikymo praktikomis.
2. **Socialinės žiniasklaidos analizė.** Nagrinėjamos žinutės ESG tema, susijusios su mados pramonės įmonėmis, siekiant įvertinti skleidžiamos informacijos nuotaikas.
3. **Kiekybiniai metodai.** Atliekama sentimentų analizė taikant teksto analizės įrankius ir mašininio mokymo algoritmus.
4. **Technologinės priemonės.** *Python* programavimo kalba teksto analizei atlikti ir *Microsoft Power BI* įrankis duomenų vizualizavimui bei vartotojų sąsajai kurti.

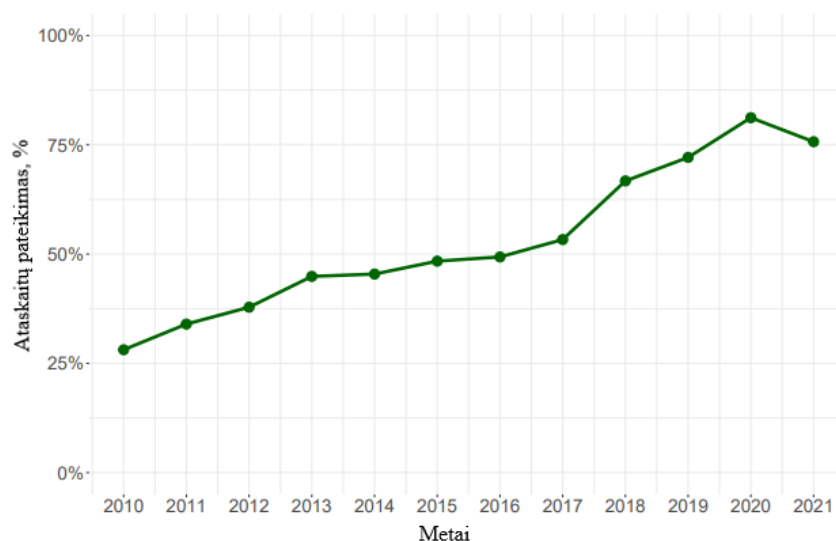
1. Literatūros apžvalga

1.1. Aplinkosaugos, socialinės ir valdymo informacijos atskleidimas ir analizė

1.1.1. ESG atsiradimas ir termino samprata

2000 m. pradžioje atsirado ESG terminas su dabartine sistema ir itin stipriai išpopuliarėjo per pastaruosius penkerius metus. C. Stedman'as [11] pabrėžia, kad didžiausios pasaulio kompanijos turi ESG programas ir yra pateikę tvarumo ataskaitas.

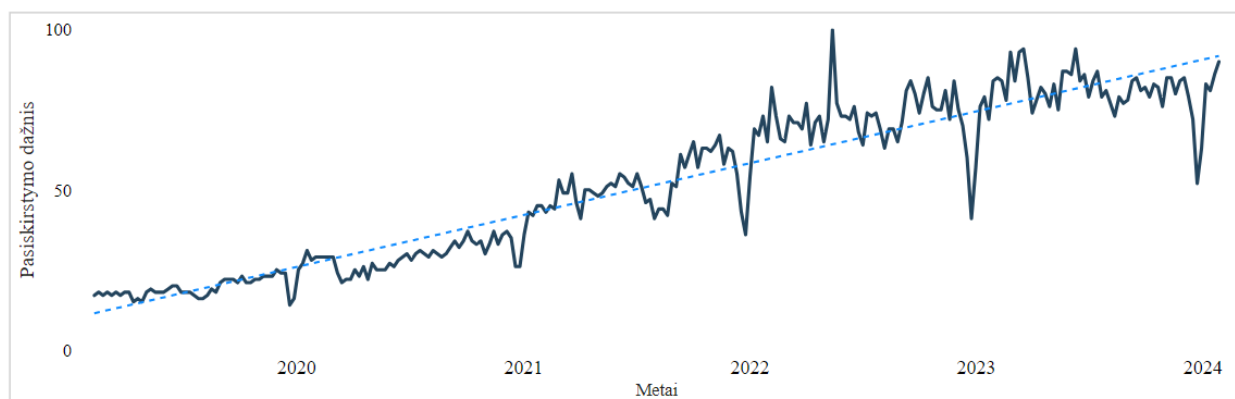
ESG ataskaitų pateikimo populiarumas pastebimas ir E. Rouen'as [12] aptartoje ESG ataskaitų pateikimo dinamikoje (žr. 1 pav.). Per dešimtmetį nuo 2010 m. išryškėjo ESG ataskaitų pateikimo augimo dinamika. Ir nors pastebimas 2021 m. sumažėjimas, tačiau taip yra tik dėl to, kad duomenų surinkimo metu ne visos įmonės, kurios teikė ataskaitas 2020 m., buvo pristatę 2021 m. ataskaitas.



1 pav. ESG ataskaitų pateikimo dinamika S&P 500¹ kompanijų [12]

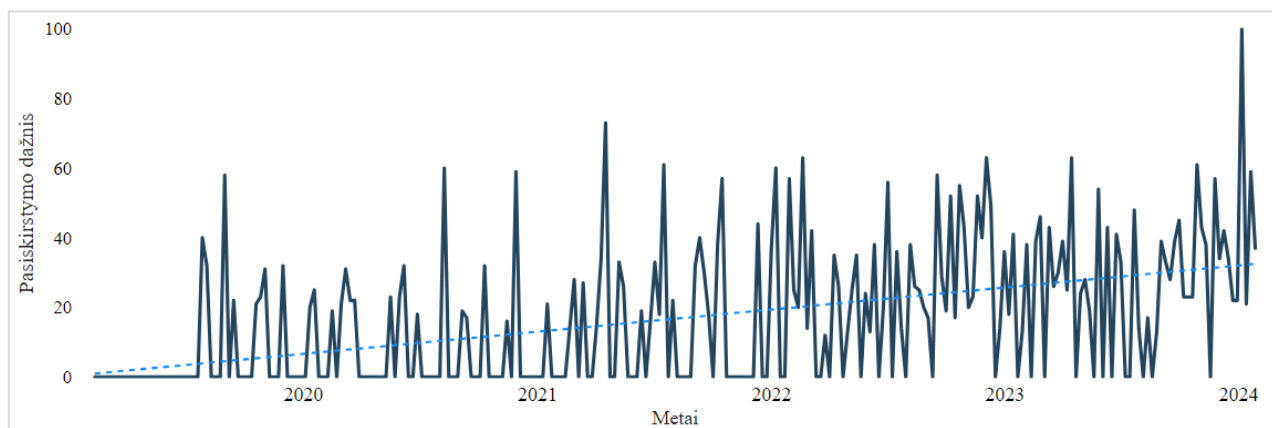
Susidomėjimą ESG tema patvirtina ir *Google Trends* [13]. Per pastaruosius penkerius metus žiniatinklio paieškoje (angl. *Web Search*) pastebimas ESG frazės paieškos augimas stebint pasauliniu mastu visose paieškos kategorijose. Stebėjimų kiekio procentinė dalis pavaizduota 2 pav., kur 100 proc. žymi viso laikotarpio didžiausią susidomėjimą.

¹ „Standard and Poor's 500“ – vertybinių popierių rinkos indeksas, rodantis 500 didžiausių JAV biržose kotiruojamų bendrovių akcijų rezultatus.



2 pav. Google Trends: susidomėjimas ESG tema pasaulyje [13]

Tuo tarpu Lietuvoje šio termino vartojimas neturi augančio populiarumo (žr. 3 pav.). Todėl tyrimui pasirinkta analizuoti visuomenės nuomonę apie didžiausias pasaulio mados pramonės įmones (angl. *Fashion & Apparel*).



3 pav. Google Trends: susidomėjimas ESG tema Lietuvoje [13]

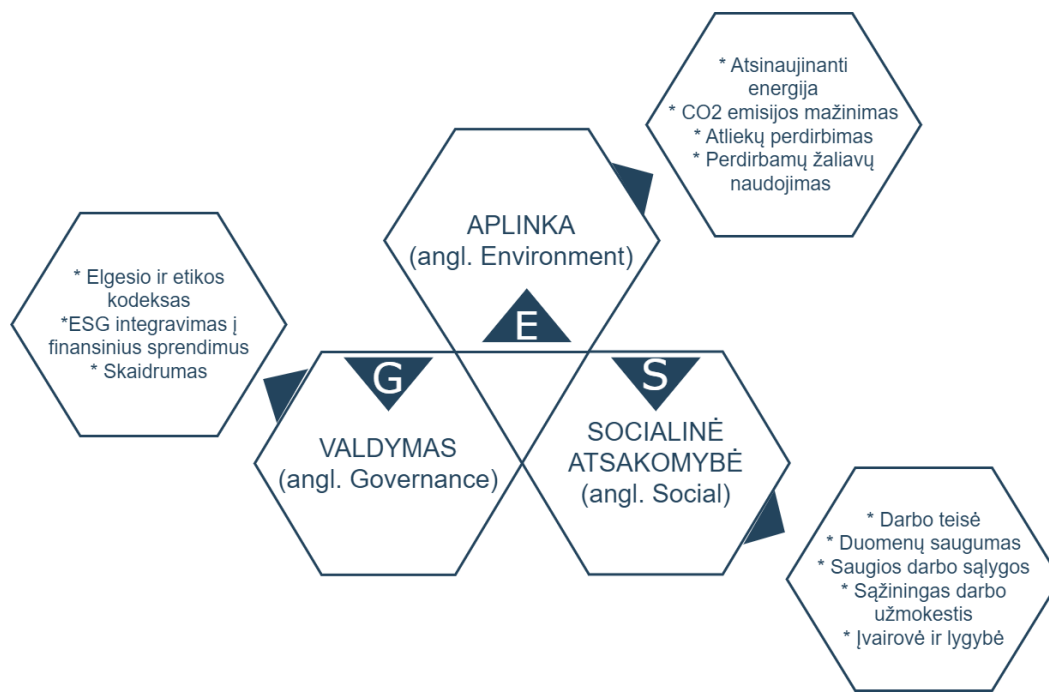
Istoriškai verslo atsakomybės idėja, kad įmonės atsakingos savo akcininkams ir visuomenei, siekia 1960-1970 m. Tuo metu susiformavo CSR (angl. *Corporate Social Responsibility* – įmonių socialinės atsakomybės) idėja [3]. Nors Jungtinės Tautos (JT) diskusijose apie ESG nedalyvauja, tačiau jos atliko lemiamą vaidmenį kuriant šį terminą ir skatinant jo sklaidą [4]. JT skatino ir rėmė daug iniciatyvų, susijusių su pasaulio ekonomika, vystymusi, aplinkosauga, žmogaus teisėmis ir kitais klausimais, turinčiais įtakos verslui ir rinkoms [5].

Labiausiai JT atsivėrė verslo sektoriui dešimtajame dešimtmetyje, kuris apibūdinamas kaip pokytis „180 laipsnių kampu“ [6]. Šiuo metu tuometinis JT generalinis sekretorius K. Annan’as padėjo pagrindus iniciatyvai, pagal kurią buvo sukurtas ESG terminas [7]. K. Annan’as paragino, kad visame pasaulyje būtų laikomasi minimalių vienodų standartų, be kurių „pasaulio ekonomika bus trapi ir pažeidžiama“. Jis paragino įmones ir verslo asociacijas „priimti, remti ir įgyvendinti pagrindines vertybes žmogaus teisių, darbo standartų ir aplinkosaugos praktikos srityse“ [8].

2004 m. Jungtinių Tautų nurodymu finansų įstaigos (angl. *IFC – International Finance Corporation*) išleido leidinį „Kam rūpi, tas laimi“ (angl. „*Who cares Wins*“), kuriame populiarinama ESG sąvoka ir rekomenduojama ESG klausimus integruoti į vertybinių popierių maklerių veiklą [9]. Ir nors

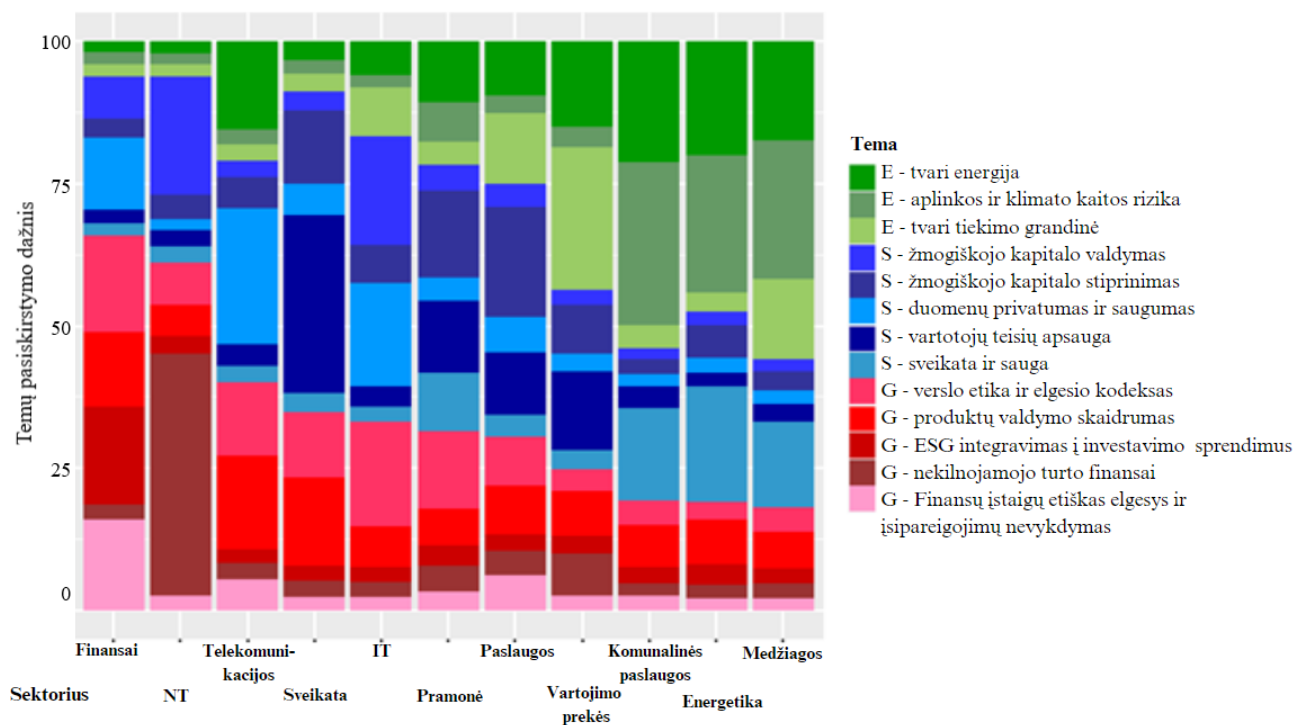
dokumento paskelbimo metais ESG terminas buvo paminėtas mažiau nei 1% diskusijų dėl pelno, 2019 m. šis rodiklis siekė 5%, o 2021 m. buvo minimas beveik penktadalyje pokalbių dėl pelno, t.y. apie 19 % [10].

Norint geriau suprasti ESG prasmę, tikslinga aptarti ką reiškia šis trumpinys. ESG – tai aplinkos, socialinės atsakomybės ir valdymo sričių derinys, kurios apima detalesnes sferas pateiktas 4 pav.



4 pav. ESG pagrindinės sritys

M. Mandas'as, et al. [14] išanalizavę ESG rizikos reitingus pateiktus *Sustainalytics* duomenų bazėje, sudarė ESG terminų žodyną, kuriuo pasinaudojant bus atlikta duomenų paieška šio darbo tyrimo dalyje (žr. 1 priedą). Taip pat atliktas tyrimas parodė, kad skirtinguose sektoriuose nagrinėjamos ESG sritys taip pat skiriasi (žr. 5 pav.), (E – Žalia, S – Mėlyna, G – Raudona)



5 pav. ESG temų pasiskirstymas pagal sektorius [14]

E. Pollman manymu ESG yra viena ryškiausių pastarųjų dviejų dešimtmečių įmonių valdymo, vadovavimo ir investavimo tendencijų [15].

B. Lutkevich'io teigimu, [16] kalbant apie ESG, daugiausiai dėmesio skiriama rizikai ir galimybėms, susijusioms su E (Aplinkosauga – angl. *Environment*) ir S (Socialiniu – angl. *Social*) komponentais. Tokios temos, kaip klimato kaita, tvarumas ir įvairovė, teisingumas ir įtrauktis patenka į šias kategorijas. Jei įmonė kalba apie anglies dvideginio išmetimo mažinimą, tai yra *E*. Kai kalbama apie etiškas tiekimo grandines, tai yra *S*. Šios dvi kategorijos yra plataus pobūdžio ir suprantamos ne tik verslo pasaulyje. ESG sąvokoje ne mažiau svarbi yra *G* (Valdymas – angl. *Governance*). Valdybos sudėtis, įmonės skaidrumas, teisių ir pareigų paskirstymas vadovybei, akcininkams ir suinteresuotosioms šalims – visa tai yra įmonės valdymo dalis. Sąžiningumas, skaidrumas, atsakomybė, teisingumas ir rizikos valdymas yra pagrindiniai valdymo principai [16].

Kaip vieną iš geros praktikos pavyzdžių B. Lutkevich'ius pateikia „Patagonia“², kuris stengiasi mažinti savo verslo ir visos tiekimo grandinės poveikį aplinkai [16]. Kompanija pereina prie 100 proc. atsinaujinančios energijos naudojimo visame pasaulyje valdomose parduotuvėse, biuruose ir platinimo centruose. „Patagonia“ perėjo prie ekologiškai užaugintos medvilnės naudojimo savo gaminiuose. Be to, „Patagonia“ yra viena iš pažangių koalicijų įkūrėjų, siekiančių pakeisti pramonę, daugiausia dėmesio skirdama darbo vietų gerinimui visame pasaulyje. Ši kompanija yra gero valdymo pavyzdys tiekimo grandinės, etikos ir sąžiningumo komponentų srityse, iš dalies dėl to, kad įmonės misija neapsiriboja vien teisės aktų laikymusi [29].

² „Patagonia“ – 1973 m. įkurta mažmeninės prekybos laisvalaikio drabužiais įmonė, turinti parduotuves daugiau kaip 10 pasaulio šalių ir gamyklas 16 šalių.

Visų dydžių įmonėms aplinkosaugos, socialiniai ir valdymo klausimai tapo svarbiausiais verslo aspektais. Įmonių ESG politiką ir praktiką atidžiai stebi investuotojai, darbuotojai, klientai, vyriausybės pareigūnai ir kitos suinteresuotosios šalys. Todėl veiksminga ESG strategija, paremta tvirtais valdymo procesais, tampa vis svarbesnė ilgalaikiai verslo sėkmei [30].

Nepaisant to, kad ESG šalininkai stipriai išpopuliarino ESG idėją, tačiau kritikai randa spragų. **E**, **S** ir **G** sujungus į vieną terminą kilo iššūkių, vienas kurių – tai termino lankstumas. Taip pat sudėtinga nustatyti ESG ir ekonominės veiklos priežastinį ryšį [17]. ESG didėja sudedamųjų dalių – subtemų sąrašas ir veikia daugiau nei 600 ESG reitingų organizacijų ir reitingų, o patys reitingai labai skiriasi [18].

Nors iki šiol ESG ataskaitų pateikimas yra savanoriškas, tačiau didžiosios įmonės daug investuoja į jų parengimą, kad užtikrintų tikslumą, aktualumą ir pateikimą. Ataskaitų poreikį lemia tai, kad informacinė aplinka stipriai keičiasi, atsiranda suinteresuotųjų šalių (investuotojų, institucijų, darbuotojų ir visos visuomenės) susidomėjimas, kuriems nebepakanka vien tik finansinių ataskaitų įmonių situacijai vertinti.

Tačiau nuo 2024 m. įsigalioja nauja įmonių tvarumo ataskaitų teikimo direktyva (angl. *Corporate Sustainability Reporting Directive*, toliau CSRD), kuri įpareigoja reguliariai skelbti ataskaitas apie įmonės veiklos poveikį aplinkai, socialinei, žmogaus teisių bei įmonių valdymą. Šie standartai reikalingi tam, kad informacija būtų aiški, aktuali ir palyginama, o siekinat užtikrinti patikimumą, ataskaitos bus audituojamos ir sertifikuojamos [76]. Pagrindiniai aspektai, kuriuos įmonės privalės atskleisti:

- **aplinkosauga:** klimato kaitos švelninimas, CO₂ emisija, žiedinė ekonomika, taršos mažinimas, atliekų mažinimas;
- **socialinė sritis:** darbuotojų sauga, žmogaus teisės, lygios galimybės, darbuotojų gerovė;
- **valdymas:** verslo etika, kova su korupcija, duomenų sauga, skaidri kultūra [76].

1.1.2. Aplinkosaugos, socialinės ir valdymo informacijos aktualumas mados sektoriuje

Pasak L. Dagilienės et al. [31] „mados pramonė yra vienas iš svarbiausių vartojimo prekėmis grindžiamų sektorių ir ekonomiškai svarbi Europos gamybos pramonės dalis“. Mados sektorių sudaro visa tiekimo grandinė apimanti žaliavų gamintojus ir tiekėjus, gamybos procesą, mažmeninę prekybą (prekės ženklus), vartotojus ir atliekų perdirbimą (žr. 6 pav.) [23].



6 pav. Tiekimo grandinės schema. Parengta pagal [23]

Dokumentiniame filme „Kitoks požiūris į madą“ (angl. *Fashion Reimagined*, 2022) [32] faktai apie mados pramonę verčia susimąstyti:

„Kasmet medvilnę renka apie 2,5 mln. vaikų“

„Kuo žemiau leidžiamės į tiekimo grandinę, tuo daugiau randame darbuotojų išnaudojimo“

„Gaminys prieš patenkant pirkėjui apkeliauja vidutiniškai 5 šalis“

Statistikos duomenimis nuo 2000 m. kasmet pagaminamų drabužių skaičius padvigubėjo ir 2014 m. viršijo 100 mlrd. [24]. Taigi, per pastaruosius 20 metų drabužių vartojimas sparčiai išaugo. Pasaulyje žmonės kasmet įsigyja daugiau kaip 80 mlrd. naujų drabužių. Tai yra daugiau kaip 400 proc. daugiau nei buvo įsigijama prieš du dešimtmečius [25].

2020 m. europiečiai drabužiams ir avalynei išleido 4,1 proc visų namų ūkių išlaidų. Per metus europiečiai drabužiams išleidžia vidutiniškai 490 eurų [26].

2021 m. drabužiai buvo svarbiausias elektroninės prekybos segmentas. Ir ateityje prognozuojamas el. prekybos augimas. *Statista* prognozuojamais duomenimis 2028 m. internetinės drabužių ir aprangos pramonės vertė sieks 1,35 trilijono JAV dolerių [27].

2021 m. lankomiausios svetainės, kuriose drabužiai buvo perkami internetu, buvo Amazon ir Target. Toliau rikiuojasi Asos, Zalando ir Zooplus [28].

Pasak atlikto tyrimo iš 18 000 apklaustųjų iš 20 šalių paskaičiuota, kad „nepaisant to, kad žmonės turi daug madingų drabužių, jie nenešioja bent 50 proc. savo drabužių“, o greitosios mados drabužiai laikomi beveik vienkartiniais ir tarnauja vos 7-8 dėvėjimus [24].

Pasaulinė statistika atskleidžia, kad mados industrija susiduria su dideliais aplinkos apsaugos išūkiais. H. Yu nuomone [33] „aplinkos tarša – svarbi pasaulinė problema, o mados pramonei priskiriama didelė kaltė dėl šios problemos“. Mados pramonė dvejopai susiduria su aplinkosaugos problemomis. Viena iš priežasčių yra ta, kad gamybai reikia milžiniškų energijos ir vandens kiekių [34]. Taip pat žaliavų gamyboje naudojama daug chemikalų: pradedant medvilnės auginimu (naudojami pesticidai) ir baigiant žaliavų dažymo bei apdorojimo procesais [32]. Mados pramonė sudaro apie 20 proc. pasaulio vandens taršos, o vieną kartą skalbiant poliesterio drabužius gali išsiskirti 700 000 mikroplastiko skaidulų, kurios gali patekti į maisto grandinę [35].

Kai kurios mados pramonės kompanijos yra sudariusios koalicijas, kad kartu spręstų aplinkosaugos ir socialines problemas, o tai padeda paspartinti pokyčius ir sumažinti riziką, kylančią dirbant su šiomis problemomis pavieniui. Pavyzdžiui, 22 drabužių prekių ženklai priklauso koalicijai „Nulis kenksmingų cheminių medžiagų išmetimo“, kurios tikslas – pagerinti ir išplėsti netoksiškų, tvarių cheminių medžiagų naudojimą tekstilės ir avalynės tiekimo grandinėje [24].

Kita opi problema – išmetamų atliekų kiekis. Greitoji moda išplito taip plačiai, nes ji leido kiekvienam atrodyti gražiai prieinamomis kainomis. Tačiau tam, kad pasiekti žemas kainas, naudojamos pigiausios sintetinės žaliavos ir pigi darbo jėga. Dėl didžiulės kolekcijų kaitos, daugybė drabužių išmetami visai nedėvėti arba jie tampa nebepatrauklūs po kelių dėvėjimų ir taip pat patenka į savartynus. Sintetiniai gaminiai ilgai nesuyra ir taip teršia aplinką.

Taigi matome, kad tiekimo grandinė yra ilga. Priklausomai nuo to, kurioje grandinės dalyje yra kompanija, ji turėtų išsikelti tikslus padedančius tausoti aplinką.

Drabužių, tekstilės ir avalynės pramonės šakos kartu yra labai imlios darbo jėgai. Tarptautinės darbo organizacijos duomenimis [36], tekstilės, drabužių, odos ir avalynės pramonėje visame pasaulyje dirba daugiau kaip 60 mln. darbuotojų. Tai reiškia, kad šis skaičius apima tik tekstilės ir aprangos gamybos sektoriuje dirbančius žmones. *Ellen MacArthur* fondas [37] teigia, kad visame pasaulyje drabužių pramonėje vertės kūrimo grandinėje dirba daugiau kaip 300 mln. žmonių, kurie, taip pat apima žmones, dirbančius mados pramonės dizaino, platinimo ir mažmeninės prekybos padaliniuose.

2022 m. duomenimis Europoje tekstilės ir aprangos pramonėje 143 000 įmonių dirba 1,3 mln. žmonių. Tekstilės sektoriuje dirba 537 482 žmonės, o drabužių gamybos sektoriuje – 760 637 žmonės [26].

Siekdami išlaikyti konkurencingas mažmenines kainas ir pelno maržas, mados prekių ženklai dažnai daro didelį spaudimą savo tiekėjams, kad šie kuo labiau sumažintų gamybos sąnaudas. Dėl to dažnai mažėja drabužių siuvėjų, kurie yra vieni mažiausiai apmokamų pramonės darbuotojų visame pasaulyje, atlyginimai. Dėmesys sąnaudų mažinimui ir įperkamo išlaikymui sudaro sudėtingas sąlygas tiekėjams mokėti didesnius atlyginimus, o tai prisideda prie nuolatinės mažų atlyginimų problemos greitosios mados sektoriuje ir dažnu atveju iššaukia darbuotojų protestus [38].

Iš paminėtų faktų galime matyti, kad mados sektorius pasaulyje susiduria ir su dideliais socialiniais išūkiais.

Kiekviena pramonės šaka stengiasi būti sėkminga savo nišoje. Tačiau sėkmę galima pasiekti tik tada, jei tenkinami trys kriterijai: ekonominis, socialinis ir aplinkosauginis. Šie kriterijai vertinami įvairiai, tačiau vienas iš vertinimo pagrindų gali būti tekstilės pramonės sertifikatai. Sertifikavimas užtikrina,

kad būtų laikomasi tinkamų standartų, didina klientų pasiekiamumą ir įvertina kokybės valdymo sistemą [39].

Mados pramonėje yra nemažai standartų, kurių siekia įmonės. Atitikimas standartams – tai geras žingsnis parodyti visuomenei, kad ji rūpinasi aplinkosauga, socialine atsakomybe ar yra draugiška gyvūnams. Standartų gali būti labai įvairių, tačiau juos galima grupuoti pagal poveikį gamtai, socialines nuostatas ir pan. Pagrindinių standartų sąrašas pateikiamas 7 pav.



7 pav. Sertifikatai pagal sritis. Parengta pagal [39, 40]

Taigi, tiek ESG ataskaitos, tiek sertifikatai leidžia įmonėms pereiti į kitą lygmenį – atsakingai ir skaidriai vykdyti verslą bei įsipareigoti vartotojams. Bendrąją prasmę sertifikatus galima laikyti ESG praktika. ESG praktika vis dažniau akcentuojama visose pramonės šakose, įskaitant mados pramonę, ir pastaruoju metu ji buvo akcentuojama tiek investicijų, tiek gamybos srityse. „Ekologiškos ir etiškos įmonės, kurios prisideda prie visuomeninės veiklos ir kurios, naudodamos ESG valdymą, užsitikrino tvaraus augimo variklius taip pat turės didelę investicinę vertę ir stabilumą“ – teigia H. Yu [33].

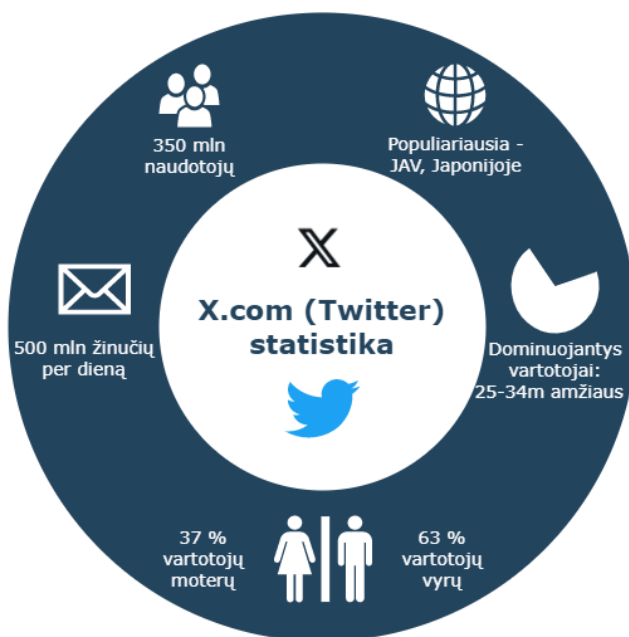
ESG praktika tampa geriausia strategija, kurią mados įmonės gali naudoti savo reputacijai gerinti. Mokslininkai nustatė, kad aukšta prekės ženklo reputacija ir vertė skatina teigiamą vartotojų reakciją, pavyzdžiui, ketinimą pirkti, pasitenkinimą ir lojalumą [41, 42]. Prekės ženklo įvaizdis ir reputacija ypač svarbūs mados pramonėje, kurią daugiausia lemia vartotojų skleidžiama informacija iš lūpų į lūpas. Reputacija daro tiesioginį poveikį įmonės veiklos rezultatams mados pramonėje labiau nei kitose pramonės šakose [43].

1.1.3. ESG tema socialinėje žiniasklaidoje

Pastaruoju metu sukuriami milijardai tekstinio turinio: žinutės, tokios kaip *WhatsApp*, *Messenger* taip pat socialiniuose tinkluose *Facebook*, *Instagram*, *Twitter (dabar X.com)*, *Google* paieškoje ir pan. Kiekvieną minutę šie šaltiniai generuoja didžiules apimtis teksto [48]. Gebant tinkamai išnaudoti informaciją, šie informacijos šaltiniai gali labai daug atskleisti apie klientų pasitenkinimą kompanijos kuriamais produktais bei vykdomomis strategijomis.

Twitter yra vienas populiariausių sklaidos kanalų sekti ir suprasti pramonės tendencijas, akcijų ir kriptovaliutų rinkos tikslus ir puikus duomenų šaltinis sentimentų analizei [19]. *Statista* duomenimis,

Twitter yra septinta pagal populiarumą socialinės žiniasklaidos platforma pasaulyje, o pagal mėnesio aktyvių naudotojų skaičių užima 14 vietą [20]. Šioje platformoje sugeneruojama 500 mln. žinučių per dieną. 2023 m. sausio mėn. duomenimis *Twitter* turėjo virš 350 mln. naudotojų, labiausiai platforma paplitusi JAV ir Japonijoje (toliau India, Brazilija, Indonezija, UK, Turkija), o didžioji auditorijos dalis – 25–34 metų amžiaus (38,5 %). Surinkti statistiniai duomenys rodo, kad didžioji dalis žmonių šiame socialiniame tinkle siekia gauti žinių, seka konkrečius prekių ženklus ar įmones ar tiesiog naršo pramoginiams tikslais. *Statista* duomenimis didžiąją auditorijos dalį sudaro vyrai (63 %) ir tik 37 % – moterys, kas yra didžiulis atotrūkis. Pagrindiniai statistiniai duomenys apie *Twitter* pateikti 8 pav.



8 pav. *Twitter* vartotojų pasiskirstymas 2023 m. duomenimis

Socialinės žiniasklaidos nuotaikos ar nuomonės yra aktualiausia ir išsamiausia informacija dėl socialinės žiniasklaidos paplitimo ir lengvo žinučių publikavimo [21]. Dėl dominuojančio jauno vartotojų amžiaus taip pat galime daryti prielaidą, kad jie yra drąsūs reikšti savo nuomonę ir aktyviai diskutuoti. Kaip teigia I. Anger et al. savo tyrime [22] tik dalis šio socialinio tinklo paskyrų yra aktyvios ir sulaukiamas mažas įsitraukimas į diskusijas. Nedidelį procentą sudaro įtakingieji (influenceriai) arba alfa naudotojų grupė, kuri sugeneruoja didelę žinučių sklaidą. O „kuo toliau skleidžiamas turinys, tuo toliau pasiekiamas naudotojo įtaka“, – teigia I. Anger.

Twitter laikomas pagrindiniu sentimentų šaltiniu, kai beveik kiekvienas entuziastingas ar socialus žmogus yra linkęs išreikšti savo nuomonę žinutėmis ar komentarais. Socialinėje žiniasklaidoje aptinkamas tekstas yra nestruktūruotas, pilnas simbolių, žodžių trumpinių, emotikonų bei žargoninių žodžių. Norint analizuoti tokį tekstą, svarbu atlikti pirminius duomenų tvarkymo ir apdorojimo etapus.

Z. Drus'as. ir H. Khalid'a savo tyrime [44] surinko ir susistemino 24 atliktus sentimentų analizės tyrimus iš socialinės žiniasklaidos duomenų atliktus 2014–2019 m. laikotarpyje. Vėliau tyrimus atliko K. H. Manguri et al. [77], M. L. Loureiro et al. [2], J. C. Lyu et al. [78] ir kiti. Keletas iš atliktų tyrimų paminėti 1 lentelėje.

1 lentelė. Sentimentų analizės tyrimai iš socialinės žiniasklaidos duomenų [2, 77, 78]

Autorius	Pavadinimas	Metodas / Įrankiai	Rezultatas	Šaltinis
Yuliyanti, Djatna ir Sukoco (2017)	Sentimentų analizė bendruomenės plėtros programų kūrime remiantis socialinės žiniasklaidos duomenimis	Leksikonu grįstas ir mašininis mokymasis	Sėkmės lygis bendruomenės plėtros programoje	Twitter
Martin-Domingo, Martin ir Mandsberg (2019)	Socialinė žiniasklaida kaip sentimentų analizės šaltinis Oro uostų paslaugų tyrimui	Mašininis mokymasis	Oro uostų paslaugų kokybės analizė	Twitter paskyra
Joyce ir Deng (2017)	2016 m. JAV prezidento rinkimų metu paskelbtų tviterio žinučių nuotaikų analizė	Leksikonu grįstas ir mašininis mokymasis	Apskaičiuotų nuotaikų ir apklausų duomenų koreliacija	Twitter
Shayaa, Wai, Chung, Sulaiman, Jaafar ir Zakaria. (2017)	Socialinės žiniasklaidos nuotaikų analizė apie įdarbinimą Malaizijoje	Leksikonu grįstas metodas	Neigiamų nuotaikų vertinimas apie įdarbinimą	Įvairūs socialinės žiniasklaidos kanalai
ElRahman, AlOtaibi & AlShehri (2019)	Twitter sentimentų analizė	Mašininis mokymasis	Populiarumas tarp KFC ir Mcdonalds	Twitter
Poecze, Ebster, Strauss & Christine (2018)	Socialinės žiniasklaidos rodikliai ir sentimentų analizė, siekiant įvertinti žinučių veiksmingumą	Mašininis mokymasis	Optimizuoti prekės ženklų komunikaciją ir suprasti vartotojų atsiliepimus	Facebook
Manguri, Amin, Ramadhan (2020)	Twitter nuotaikų analizė dėl pasaulinių COVID-19 protrūkių	Leksikonu grįstas ir mašininis mokymasis	Ištyrta emocijos, kai kalbama apie Covid-19 temą	Twitter
Loureilo, Allo, Coello (2022)	Karšta Twiteryje: gaisrų emocinio poveikio vertinimas naudojant nuotaikų analizę	Vader sentimentų analizė	Reakcija į oro taršą	Twitter ir oro taršos duomenys
Lyu, Han, Luli (2021)	Su COVID-19 vakcina susijusios diskusijos Twiteryje	Leksikonu grįstas metodas	COVID-19 žinučių poliariškumo tyrimas	Twitter

Sentimentų analizė yra labai plačiai nagrinėjama sritis. *Twitter* platforma taip pat dažnai minima kaip vienas tinkamiausių duomenų šaltinių. Ir nors ESG tema vis dažniau tampa analizės objektu tarp įmonių, kol kas mokslinių straipsnių, susijusių su ESG tema ir sentimentų analize nepavyko rasti.

Tolesnėse šio projekto dalyse pasitelkiant NLP metodikas ir matematinius modelius bus analizuojamos *Twitter* žinutės siekiant išsiaiškinti kokią nuomonę apie mados pramonės ESG praktiką skleidžia suinteresuotieji asmenys, kokiomis temomis dažniausiai kalba ir kaip tai gali paveikti įmonių reputaciją.

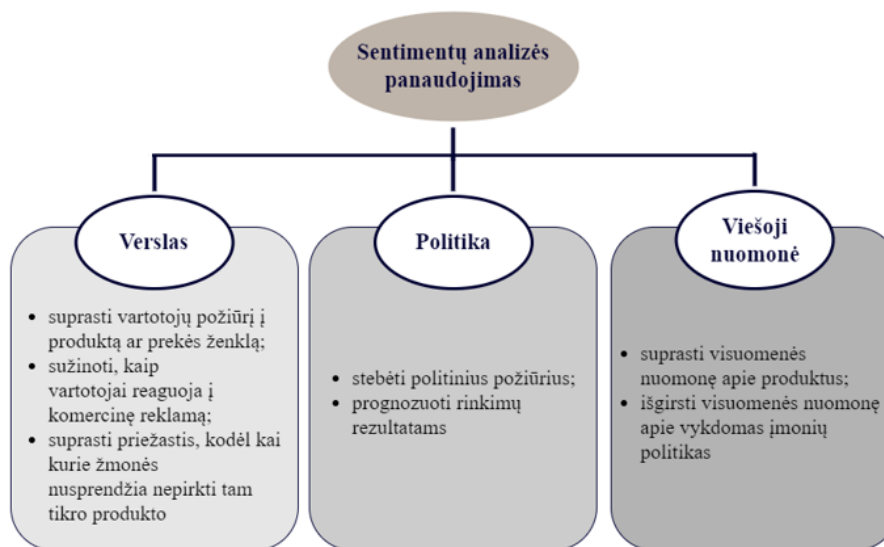
1.2. Sentimentų analizė

1.2.1. Sentimentų analizės pagrindiniai aspektai

M. D. Devika et al. sentimentų analizę (SA), dar vadinamą nuotaikų analize (angl. *Opinion mining*), apibūdina kaip intelektualų procesą, kurio metu išgaunami vartotojo jausmai ir emocijos. Tai viena iš natūralios kalbos apdorojimo (NLP) sričių [45]. Z. Drus'as et al. teigia, kad sentimentų analizė apima teksto klasifikavimą į įvairius jausmus, tokius kaip teigiamas ar neigiamas, laimingas, liūdnas ar neutralus ir t.t. Taigi, galutinis sentimentų analizės tikslas yra iššifruoti pagrindinę nuotaiką, emocijas ar jausmus [44]. R. S. Jagdale pabrėžia, kad sentimentų analizė naudojama siekiant įvertinti kalbėtojo, rašytojo ar kito subjekto jausmus, susijusius su tam tikra tema, arba bendrą konkretaus įvykio, pokalbio, forumo, veiksmo ar failo orientaciją ir t. t. [46].

Natūralios kalbos apdorojimas (toliau NLP), sujungia kompiuterinę lingvistiką – taisyklėmis pagrįstą žmogaus kalbos modeliavimą su statistiniais ir mašininio mokymosi modeliais, kad kompiuteriai ir skaitmeniniai prietaisai galėtų atpažinti, suprasti ir kurti tekstą bei kalbą [47]. NLP – tai dirbtinio intelekto (DI) šaka, kuri atpažįsta žodinį ir rašytinį tekstą ir gali apibendrinti didelius teksto kiekius, bei įvertinti teksto ar kalbos ketinimus ar nuotaikas [48].

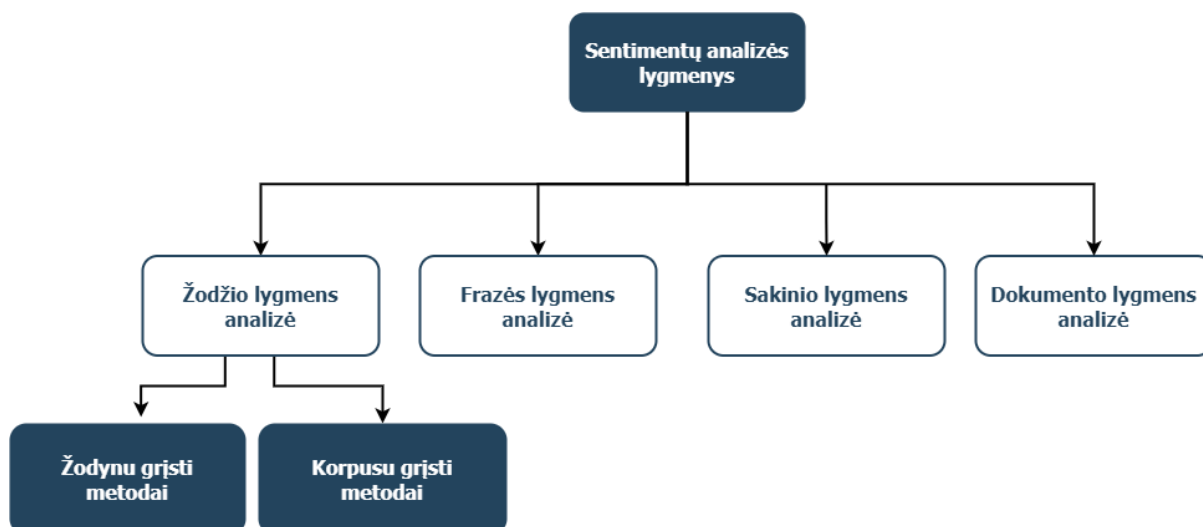
Spartus internetinių programų, tokių kaip socialinės žiniasklaidos platformos ir tinklaraščiai, augimas paskatino analizuoti komentarus ir apžvalgas savo kasdienėje veikloje. Daugelis vartotojų naudoja socialinę erdvę, norėdami išreikšti savo nuomonę [49]. Žmonių nuomonės gali būti naudingos įmonėms, vyriausybėms ir atskiriems asmenims renkant informaciją ir priimant nuomonėmis pagrįstus sprendimus. Sentimentų (kitais nuotaikų / nuomonių) analizė pastaraisiais metais sulaukė didelio pripažinimo ne tik tarp mokslininkų, bet ir tarp įmonių, vyriausybių ir organizacijų (žr. 9 pav.) [50]



9 pav. Sentimentų analizės panaudojimas. Parengta pagal [49, 51]

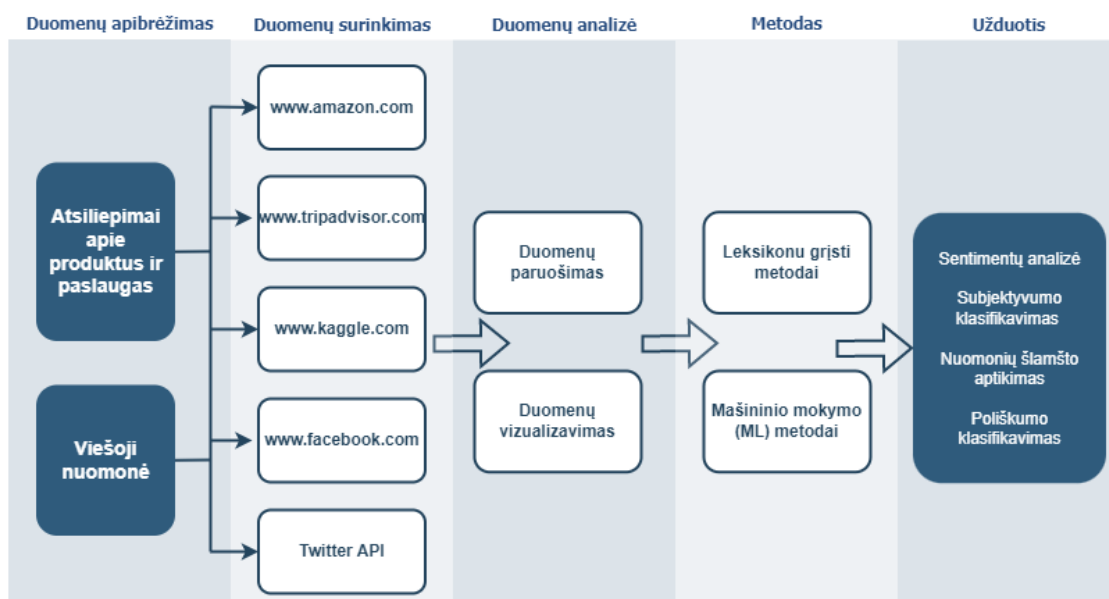
Sentimentų analizėje daugiausia dėmesio skiriama teksto poliariskumui (teigiamas, neigiamas, neutralus), tačiau ji taip pat neapsiriboja tik poliariskumu, o padeda nustatyti konkrečius jausmus ir emocijas (piktas, laimingas, liūdnas ir t. t.), skubumą (skubus, neskubus) ir net ketinimus (domina

arba nedomina) [52]. P. Sudhir'as [48], M. Wankhade et al. [49], N. Singh'as [51] savo tyrimuose sentimentų analizę išskiria į lygmenis kaip pavaizduota 10 pav.



10 pav. Sentimentų analizės lygmenys [49, 51, 52]

- **Dokumento lygmens sentimentų analizė.** Šis metodas gali būti tinkamas klasifikuoti knygų skyriams, siekiant suprasti, ar knyga turi yra teigiamą požiūrį, neigiamą ar neutralų [49].
- **Sakinių lygmens sentimentų analizė.** Šis metodas taikomas, kai dokumente pasitaiko skirtingų nuotaikų. Kiekvieno sakinio poliariškumas bus nustatomas atskirai, taikant tas pačias metodikas kaip ir dokumento lygmenyje, tačiau naudojant daugiau mokymo duomenų ir apdorojimo išteklių [49]. Šio lygmens analizę apsunkina darbas su sąlyginiais sakiniais ir dviprasmiškais teiginiais [53].
- **Frazių (aspekto) lygmens sentimentų analizė.** Šiuo būdu atliekama nuotaikų analizė kurios metu nuomonės žodžiai bus išgaunami frazių lygmeniu ir klasifikuojami. Kiekvienoje frazėje gali būti keli arba vienas aspektas. Kaip teigia B. Liu [64], dažnai minimi daiktavardžiai tikėtina, kad tai yra tikrieji aspektai. Analizuodami tekstą norime suprasti, kokius konkrečius aspektus ar savybes žmonės mini teigiamai, neutraliai ar neigiamai [49]. Aspektais pagrįsta nuotaikų analizė gali padėti organizacijoms automatiškai rūšiuoti ir analizuoti klientų duomenis, automatizuoti procesus, pavyzdžiui, klientų aptarnavimo užduotis, leidžia įgyti daug įžvalgų [48]. „Šio fotoaparato akumuliatoriaus veikimo laikas yra per trumpas“, aspektu pagrįstas klasifikatorius galėtų nustatyti, kad šiuo sakiniu išreiškiama neigiama nuomonė apie aptariamo gaminio akumuliatoriaus veikimo laiką [53].
- **Žodžių lygmens sentimentų analizė** – tai populiarus ir sėkmingas nuomonės tyrimo metodas. Pavyzdžiui, „puikus“, „nuostabus“ – teigiamas požiūris. Yra daug duomenų bazių, kuriose pateikiama informacija apie būdvardį ir klasę, kuriai jis priklauso. Sinonimas parodo, kaip kas nors jaučiasi dėl ko nors gero, o antonimas – kaip kas nors jaučiasi dėl ko nors blogo [51].



11 pav. Bendrinė sentimentų analizės atlikimo eiga [49]

M. Wankhade [49] pateikia bendrinę sentimentų analizės procedūrą, kuri atvaizduota 11 pav. Naudojantis šia schema, atliekama *Twitter* duomenų analizė, siekiant išsiaiškinti visuomenės nuomonę ESG tema.

1.2.2. Sentimentų analizės tyrimo problematika

Taikant sentimentų analizės metodą susiduriame su įvairiais iššūkiais, pradedant teksto apdorojimu ir baigiant skaičiavimo sąnaudomis [49]. Itin daug iššūkių pasitaiko nagrinėjant nestruktūruotą tekstą, kaip šiuo atveju – *Twitter* žinutes. Žinučių rašymo stilius yra neformalus, rašytojo nevaržo jokios taisyklės ir apribojimai. K. A. Sharou et al. [66] nagrinėja „triukšmą“, kaip vieną iš apsunkinančių sąlygų teksto analizei. Jo teigimu, nors išankstinis duomenų apdorojimas yra neabejotinai svarbus NLP, tačiau platesnis supratimas apie įvairius triukšmo šaltinius ir kaip su jais elgtis – tai aspektas, kuris iš esmės nebuvo nagrinėtas.

Triukšmo tipai NLP gali būti:

- **gramatinės klaidos.** Šis triukšmo tipas reiškia nukrypimą nuo kalbos gramatikos taisyklių, išskyrus rašybos klaidas: žodžių rašyba, žodžių suderinamumas, žodžių eiliškumas;
- **skyrybos ženklai ir jų pakartojimai.** Neteisingas skyrybos ženklų vartojimas gali pakeisti sakinio prasmę. Pasikartojantys skyrybos ženklai prideda emocijas. Pvz. „Really! You want me to go now????“, „What???? This is really annoying!!!“;
- **ortografija**³. Šis triukšmo tipas susijęs su žodžių rašymo būdu. Tai rašybos klaidos, žodžių slėpimas, žodžių ilginimas, santrumpos, simboliai. Žodžių slėpimas arba maskavimas dažnai skirtas smurto arba keiksmažodžių maskavimui, kaip pavyzdžiui „ki11“ vietoj „kill“ [68]. Žodžių ilginimas pridedant papildomą pasikartojančią raidę, pvz. „Yes, Nooow!“. Simboliai arba emotikonai tekstui suteikia papildomos prasmės pridedant emocijas: šypsena 😊, nuostabą 😲, pyktį 😡 ir pan. [66];

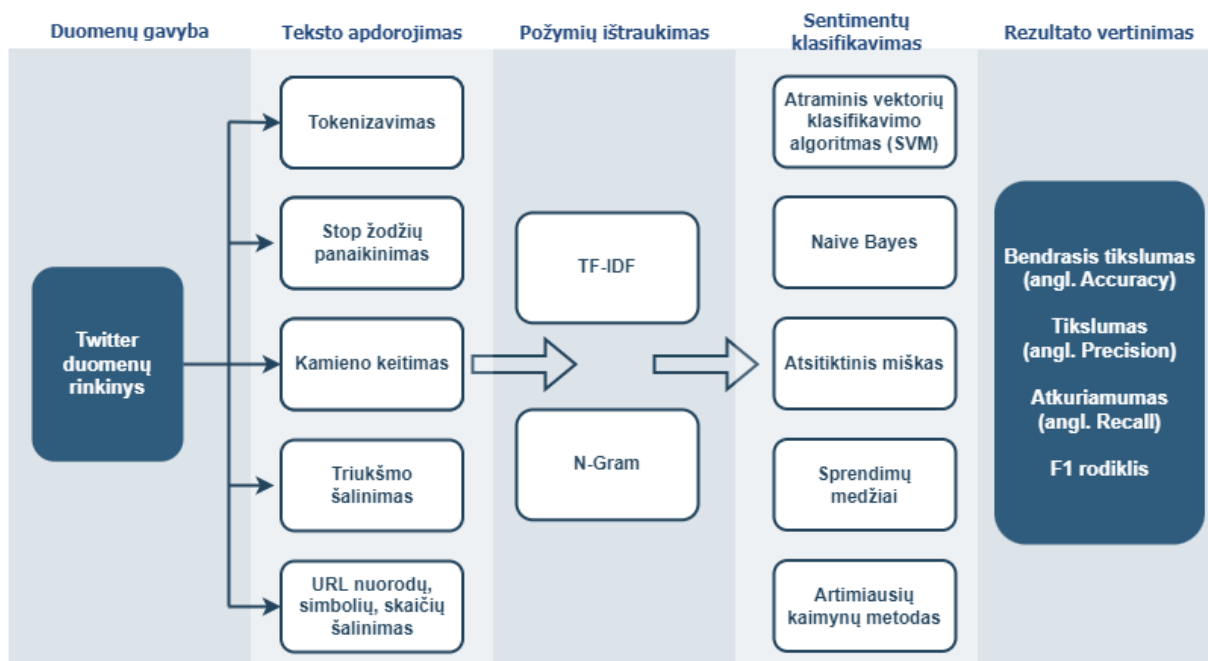
³ Ortografija – tai žodžių ir jų formų rašymo taisyklių visuma. (<https://vlkk.lt/>)

- **internetinis žargonas arba žodžių trumpiniai ir pakaitalai.** Dažnu atveju visuotinai paplitę trumpiniai „Professor“ trumpinys „Prof“, „LOL“ trumpinys „Laugh Out Loud“, „tom“ – „tomorrow“, „msg“ – „message“, „2day“ – „today“, „4you“ – „for you“ [67];
- **neformalios žodžių formos arba šnekamosios kalbos vartojimas.** Šis tipas – tai atvejai, kai keli žodžiai rašomi kartu kaip vienas žodis, laikantis tam tikrų taisyklių [68], pvz. „want to“ → „wanna“, „I am“ – „I’m“;
- **pauzę užpildantys žodžiai:** „uh“, „er“, „um“. Patys savaime jie neturi jokios reikšmės, tačiau gali rodyti svarbius aspektus, pvz., dvejonę ar netikėtą reakciją (teigiamą ar neigiamą);
- **kelių kalbų naudojimas.** Pavyzdžiui, junginio panaudojimas – „努力ing“ tai dviejų kalbų junginys. Kinų kalba reiškia „sunkiai dirbti“ ir pridėta angliška priesaga -ing. „This is so cute“ – „This is so kawaii“ [67, 68].

1.3. Teksto apdorojimas ir sentimentų analizės metodai

Socialinės žiniasklaidos nuotaikų analizei NLP tapo esminiu verslo įrankiu, leidžiančiu atskleisti paslėptų duomenų įžvalgą iš socialinės žiniasklaidos kanalų. Sentimentų analizė gali analizuoti kalbą, vartojamą socialinės žiniasklaidos įrašuose, atsakymuose, apžvalgose ir kt., siekiant išgauti požiūrį ir emocijas, reaguojant į produktus, reklamas ir įvykius – informaciją įmonės gali naudoti kurdamas produktus, reklamines kampanijas ir kt. Kaip jau minėta anksčiau, sentimentų analizės tyrimai aktualūs tiek versle, politikoje ar siekiant sužinoti viešąją nuomonę socialinėje erdvėje.

Sentimentų analizė paprastai veikia naudojant natūralios kalbos apdorojimo (NLP) metodus, kad būtų galima analizuoti ir suprasti tekste išreikštą nuotaiką.



12 pav. Sentimentų analizės procesas [59]

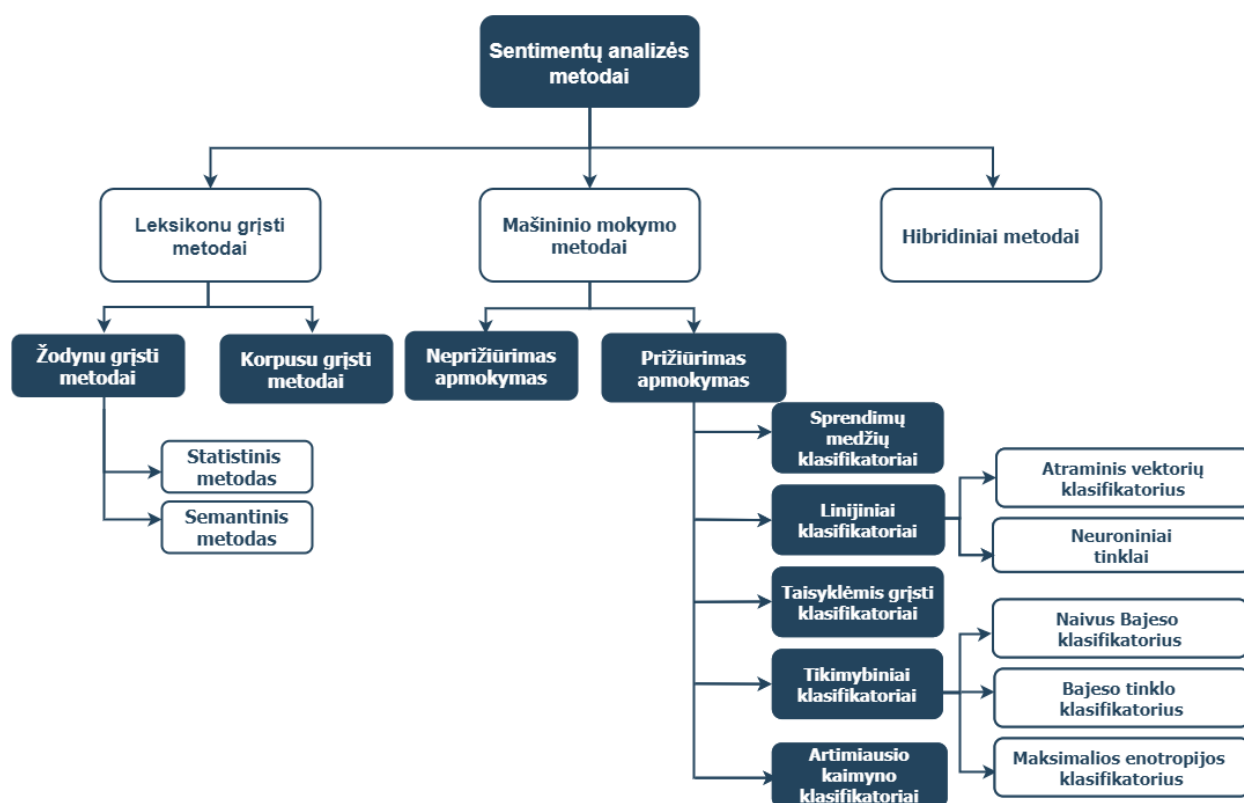
Procesas apima kelis etapus [58, 59], kurie atvaizduoti 12 pav. Metodologijos dalyje detaliau apžvelgiamas kiekvienas iš šių etapų, o tyrimo dalyje atliekamas praktinis teksto apdorojimas, algoritmo parinkimas ir įvertinimas.

P. Sudhir'as [48], M. Wankhade et al. [49], W. Medhat'as [57] straipsniuose pateikia algoritmus ir metodus sentimentų analizei nagrinėti, metodų privalumus ir trūkumus, bei palygina gautus rezultatus tarpusavyje (žr. 2 lentelę).

2 lentelė. Matematinų metodų tikslumas naudojant mašininio mokymo algoritmus [48]

Klasifikavimo modelio tipas	Klasifikavimo modelis	Bendras tikslumas
Įprastas mašininis mokymasis	Naivusis Bajesas (NB)	54,77 %
	Sprendimų medis (DT)	87,53 %
	Atraminų vektorių metodas (SVM)	88,30 %
	K-artimiausių kaimynų	88,86 %
Gilusis mokymasis	LSTM	88,50 %
	GRU	88,50 %
Transformerių modeliai	BERT	89,50 %
	BERT-UDA	95,22 %

Sentimentų analizei tirti naudojami skirtingi matematiniai metodai ir algoritmai pradedant leksikonu grįstais metodais (angl. *Lexicon-based*) ir baigiant mašininio mokymosi metodais. Keletas dažniausiai naudojamų metodų pateikti schemeje (žr. 13 pav.)



13 pav. Sentimentų klasifikavimo metodai. Parengta pagal [49, 57]

Sentimentų analizei atlikti galimi keli metodai:

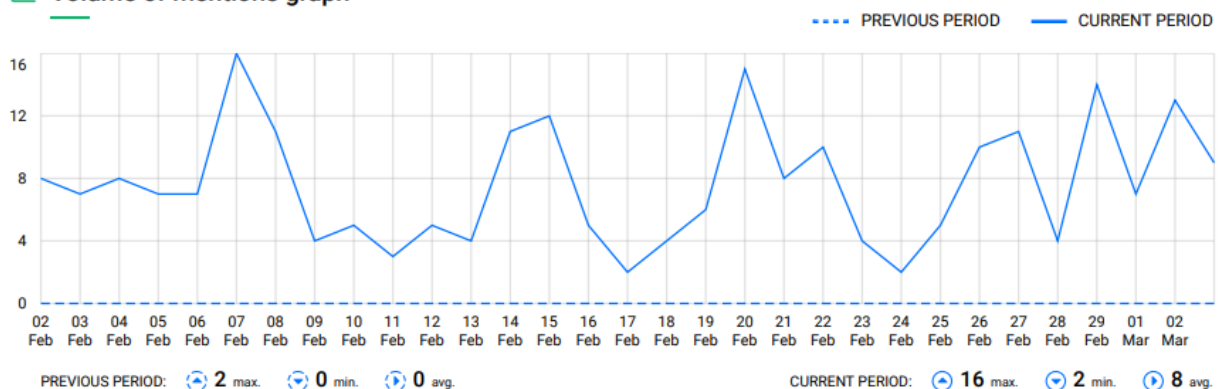
- mašininis mokymas (angl. ML);
- leksikonu grindžiamas metodas (angl. Lexicon based);

- 28

Summary of mentions



Volume of mentions graph



15 pav. Brand24 socialinės žiniasklaidos žinučių sentimentų analizės pavyzdinė švieslentė [54]

Generatyvinės kalbos DI modelis (*ChatGPT*) taip pat gerai geba spręsti sentimentų analizės užduotis.

Pateikus vieną iš nagrinėjamų *Twitter* žinučių *ChatGPT* – „**High-street #fashion store #Zara accused of using #slave #labour at factories in #Argentina**“ – grąžino ne tik atsakymą, kad tai yra neigiamo pobūdžio žinutė: „Ši *Twitter* žinutė būtų priskiriama neigiamoms nuotaikoms“, bet ir detalių paaiškinimą:

„**Neigiama kalba:** Tokių žodžių, kaip „kaltinamas“ (angl. accused) ir „vergų darbas“ (angl. slave labour), vartojimas turi stiprų neigiamą atspalvį. Šie terminai siejami su neetiška praktika ir žmogaus teisių pažeidimais, kuriuos visuomenė vertina neigiamai.

Prieštaringas klausimas: Tvityrje aptariamas rimtas ir prieštaringas klausimas, susijęs su įtariama neetiška žinomo prekės ženklo veikla. Tokios temos paprastai sukelia neigiamas nuotaikas, nes yra susijusios su socialine neteisybe ir išnaudojimu.

Poveikis prekės ženklo įvaizdžiui: Konkretaus prekės ženklo (šiuo atveju „Zara“) paminėjimas neigiamame kontekste (kaltinimas vergų darbo naudojimu) greičiausiai pakenktų prekės ženklo įvaizdžiui ir reputacijai, o tai dar labiau sustiprintų neigiamas tviterio nuotaikas.

Norint nustatyti bendrą teksto nuotaiką, atliekant nuotaikų analizę dažnai atsižvelgiama į kontekstą ir konkrečius pavartotus žodžius. Šiuo atveju kontekstas ir parinkti žodžiai aiškiai rodo neigiamą nuotaiką.“

Generatyvinės kalbos modelis geba spręsti sentimentų analizės užduotis, tačiau neapdoroja didelių duomenų kiekių. Tokiu atveju pats ChatGPT siūlo naudotis automatizuotomis priemonėmis.

Vienas iš šio projekto tikslų – sukurti vartotojo sąsają atvaizduojant sentimentų analizės rezultatus vizualioje, lengvai valdomoje švieslenteje. Tam tikslui pasiekti pasitelkta 14 pav. ir 15 pav. prototipais, o švieslentei sukurti panaudotas *Microsoft Power BI* įrankis (žr. 2 priedą).

1.5. Literatūros analizės apibendrinimas

ESG – aplinkosaugos, socialinės atsakomybės ir valdymo informacija, pastaraisiais metais tapo svarbia įmonių veiklos vertinimo dalimi. Šis požiūris skatina įmones ne tik skelbti finansinius rezultatus, bet ir demonstruoti jų atsakomybę aplinkosaugos, socialinės sąmoningumo ir etiško valdymo srityse.

Dauguma mokslininkų – C. Stedman’as [11], E. Rouen’as [12], M. Mandas’as, et al. [14] – savo straipsniuose pabrėžia ESG įvairiapusę naudą įmonėms. Nors daugiausiai akcentuojamas tvarumas ir aplinkosauga (**E**) bei darbuotojų gerovė (**S**), tačiau B. Lutkevich’ius [16] ragina nepamiršti ir valdymo (**G**) temos svarbos. Jis teigia, kad sąžiningumas, skaidrumas, atsakomybė, teisingumas ir rizikos valdymas yra pagrindiniai valdymo principai.

Mados sektorius yra ypač jautrus visuomenės nuomonei, nes vartotojų įpročiai ir požiūris tiesiogiai veikia prekių ženklų reputaciją ir pardavimus. Atsižvelgiant į didėjančią susirūpinimą aplinkos tarša, darbo sąlygomis mados pramonėje, įmonės vis dažniau skelbia apie savo ESG iniciatyvas, siekdamos ne tik patenkinti reguliavimo reikalavimus, bet ir gerinti savo įvaizdį vartotojų akyse.

Sentimentų analizė yra vienas iš efektyvių metodų, skirtų analizuoti, kaip visuomenė reaguoja į įmonių vykdomą veiklą. Jis apima tekstų, kuriuose gali būti įvairių emocijų, poliariškumo (teigiamas, neigiamas, neutralus) ir nuotaikų įvertinimą. Socialinėje žiniasklaidoje, pvz. *Twitter*, kur vartotojai laisvai išreiškia savo nuomones, sentimentų analizė leidžia greitai sužinoti visuomenės požiūrį į konkrečias temas, įskaitant ESG praktikas. *Twitter* žinučių duomenis savo tyrimams naudojo aptarti mokslininkai: M. L. Loureiro et al. [2], K. H. Manguri et al. [77], J. C. Lyu et al. [78] ir daug kitų, kuriuos aprašė Z. Drus’as ir H. Khalid’as savo tyrime [44].

Moksliniuose tyrimuose sentimentų analizei dažniausiai naudojami NLP įrankiai ir mašininio mokymosi metodai, tokie kaip Naivusis Bajesas, sprendimų medžiai, atraminių vektorių metodas (SVM), ir naujesni, pavyzdžiui, LSTM, GRU, BERT modeliai. Šie metodai padeda analizuoti didelius neapdorotų duomenų kiekius ir nustatyti dominuojančias nuotaikas socialinės žiniasklaidos tekstuose.

2. Tyrimų metodologija

Šioje darbo dalyje pateikiama darbo metodologija, skirta analizuoti *Twitter* vartotojų žinutes ESG tema, kurios sutinkamos kalbant apie mados pramonę. Metodologija apima duomenų gavybą, teksto apdorojimą, modelių analizę, metrikų vertinimą, duomenų vizualizavimą ir vartotojo sąsajos parengimą. Tyrimas realizuotas panaudojant *Python* programavimo kalbą ir *Microsoft Power BI* įrankį duomenų vizualizavimui atlikti (žr. 2 priedą).

Python programavimo kalba suteikia platų įrankių ir bibliotekų spektrą, skirtą atlikti konkrečias NLP užduotis. Daugelis jų yra natūralios kalbos įrankių rinkinyje (sutrumpintai NLTK) – atvirojo kodo bibliotekų rinkinyje, skirtame NLP programoms kurti.

NLTK apima bibliotekas užduotims, pvz., sakinių analizei, žodžių segmentavimui, kamieno sudarymui ir lematizacijai (žodžių apkarpymo iki šaknų metodai) ir atpažinimo (frazių, sakinių, pastraipų skaidymo) bibliotekas. Tai taip pat apima bibliotekas, skirtas įgyvendinti tokias galimybes kaip semantinis samprotavimas, galimybė daryti logiškas išvadas, pagrįstas faktais, ištrauktais iš teksto [47].

2.1. Duomenų gavyba

Sentimentų analizei atlikti darbe analizuojamos *Twitter* žinutės (kitai vadinamos tweet). Ir nors savo darbe vis dar naudoju seną platformos pavadinimą, tačiau nuo 2022 m. balandžio mėn. *Twitter* koncepcija pasikeitė, kuomet E. Musk'as įsigijo platformą ir ją pervadino į X. *Twitter* leido publikuoti tik trumpas – iki 140 simbolių – žinutes, tačiau X platformoje galima skelbti beveik viską, įskaitant kelių valandų trukmės vaizdo įrašus ir ilgus pranešimus [69].

Šie pasikeitimai galėjo nulemti ir privatumo politikos pasikeitimus. Anksčiau galioję duomenų ištraukimo metodai, kaip pavyzdžiui *Python* biblioteka *Tweepy* gana lengvai pritaikyta *Twitter* žinučių analizei, nebetika. 3 lentelėje pateikiami *Twitter* žinučių duomenų gavybos pokyčiai kuriuos savo tyrime susistemino D. Antonakaki [70].

3 lentelė. *Twitter* duomenų gavybos politika

Laikotarpis	Politika
2006–2010 m.	Baltasis sąrašas, sušvelninti apribojimai
2010–2012 m.	OAuth2
2012–2017 m.	Griežti API apribojimai
2017–2020 m.	Premium / Kompanijų API, nėra viešų duomenų
2020–2021 m.	APIv2
nuo 2022 iki dabar	Sustiprinta duomenų nuskaitymo politika

Siekiant surinkti tikslingą informaciją, *Twitter* paieškos lauke vedama užklausa, kurią sudaro keletas raktažodžių. Tyrimui atlikti pasirinktas mados sektorius (angl. *Fashion & Apparel*). Kaip jau buvo paminėta, mados sektorių sudaro visa tiekimo grandinė apimanti žaliavų gamintojus ir tiekėjus, gamybos procesą, mažmeninę prekybą (prekės ženklus), vartotojus ir atliekų perdirbimą (žr. 6 pav.) [23], tačiau vartotojas yra arčiausiai prekės ženklo (mažmenininko), todėl galima tikėtis, kad daugiausiai atsilepimų apie jį ir paliks. *Twitter* žinutės ieškomos pasirenkant mados industrijos stambiausių prekinių ženklų pavadinimus iš 100 didžiausių pasaulyje listinguojamų drabužių ir

mados pramonės įmonių [71]. Paieškos lauke panaudojus tokius raktažodžius kaip @Zara, @HM, @GAP, @Nike, @Adidas, @Dior, @Prada ir kt. randamos tikslinio sektoriaus žinutės.

Norint atrasti tikslią informaciją ESG tema, reikia nurodyti raktažodžius *Twitter* paieškos lauke su simboliu #. Pasitelkus literatūros apžvalgą, išanalizavus mados pramonės silpnąsias sritis bei pasitelkus sudarytais ESG žodynais, pasirinkti terminai ir frazės geriausiai atspindės ESG temą.

„E“ sritį geriausiai apibūdina šie raktažodžiai: #Carbon, #Emission, #Renewable, #Waste, #Water, #Plasticbags, #Manufacture ir kt.

„S“ – #Workforce, #Humancapital, #Labour, #Privacy, #recruitment, #Accident, #Safety, #Injury, #HumanRights, #Wages

„G“ – #Governance, #Corporate, #Stockmarkets, #Supplychain, #Shareholder, #Corruption, #Transparency, #BusinessEthics, #Investment, #Finance

Taigi, paminėtų raktažodžių pagalba, panaudotų kartu su kompanijų pavadinimais, gautos tikslinės žinutės ESG tema iš mados pramonės įmonių. Toliau žinutės yra apdorojamos panaudojant teksto apdorojimo bibliotekas.

2.2. Teksto apdorojimas

Gautoms žinutėms atliekamas **išankstinis teksto apdorojimas** (angl. *Text Preprocessing*) ir skaidymas (angl. *Tokenisation*). Teksto apdorojimo procesas – svarbus etapas, kuris gali nulemti galutinį rezultatą. Žinutės yra skaidomos į atskirus žodžius, dar vadinamus žetonais (angl. *Tokens*).

Teksto duomenys išvalomi pašalinant nesusijusią informaciją, pvz., specialiuosius simbolius, skyrybos ženklus ir stabdomuosius žodžius (angl. *stopwords*), raidės verčiamos į mažąsias. Žodžiams, turintiems panašią prasmę gali būti atliekamas kamienizavimas ir lematizavimas. Kamieno keitimas – kamienizavimas (angl. *Stemming*) – padeda suvienodinti žodžius, jei jie turi tą pačią prasmę. Lematizavimas – tai dviejų ar daugiau žodžių sujungimo į vieną žodį procesas. Teksto, kamienizavimo ir lematizavimo pavyzdys pateiktas 16 pav.

```
Text: Burberry becomes the first luxury brand to commit to being ClimatePositive by 2040  
Stem: ['burberri', 'becom', 'first', 'luxuri', 'brand', 'commit', 'climateposit', '2040']  
Lemma: ['Burberry', 'become', 'first', 'luxury', 'brand', 'commit', 'ClimatePositive', '2040']
```

16 pav. Teksto apdorojimo pavyzdys, taikant lematizavimą ir kamienizavimą

Stop žodžiai – tai dažnai pasikartojantys žodžiai, kurie žinutėms nesuteikia prasmės. Tai labiausiai paplitę žodžiai bet kurioje natūralioje kalboje. Norint analizuoti teksto duomenis ir kurti NLP modelius, šie stabdomo žodžiai neduoda pridėtinės vertės dokumento prasmei [59]. Tačiau kaip teigia M. Wankhade et al. stop žodžių pašalinimą reikia atlikti atsakingai, nes neiginiai taip pat priskirti stop žodžiams. Jų pašalinimas iš žinutės gali suteikti frazei visai kitą prasmę.

Tyrimui surinktų ir apdorotų duomenų saugojimui pasirinktas *.parquet formatas vietoj *.csv. *Apache Parquet* yra atvirojo kodo, į stulpelius orientuotas duomenų failų formatas, skirtas efektyviai saugoti ir gauti duomenis. Jame pateikiamos veiksmingos duomenų suspaudimo ir kodavimo schemas bei padidintas našumas, kad būtų galima apdoroti sudėtingus duomenis dideliais kiekiais.

Apache Parquet panašus į kitus *Hadoop* sistemoje esančius stulpelinės saugyklos failų formatus, t. y. *RCFile* ir *ORC*.

Apache Parquet efektyvesnis nei *.csv formatas. Užklauskas teikiant stulpelinėje saugykloje galima lengvai praleisti nereikšmingus duomenis. Todėl užklauskos užima mažiau laiko nei į eilutes orientuotos duomenų bazės. Toks duomenų saugojimo būdas leidžia sutaupyti įrangos resursus, pagerina nuskaitymo laiką, taigi ir bendras sąnaudas [74]. Šiuo formatu užsaugoti duomenys toliau bus panaudoti duomenų vizualizavimui *Microsoft Power BI* aplinkoje.

Požymių išskyrimas (angl. Feature extraction): iš teksto išgaunamos atitinkamos funkcijos, pvz., žodžiai, n-gramos ar net kalbos dalys.

Terminų dažnis-atvirkštinis dokumentų dažnis (TF-IDF) yra plačiai naudojamas statistinis metodas natūralios kalbos apdorojimo ir informacijos paieškos srityse. Juo matuojama termino svarba dokumente, palyginti su dokumentų rinkiniu (t. y. korpusu). Žodžiai teksto dokumente paverčiami svarbos skaičiais taikant teksto vektorizavimo procesą. Yra daug įvairių teksto vektorizavimo vertinimo schemų, TF-IDF yra viena iš labiausiai paplitusių [59].

Terminų dažnis (angl. *Term Frequency* – toliau TF) vertina, kaip dažnai terminas pasitaiko dokumente ir apskaičiuojamas kaip termino pasikartojimų dokumente skaičiaus ir visų dokumente esančių žodžių skaičiaus santykis (žr. (1) formulę):

$$tf(t, d) = \frac{f_{t,d}}{DF}; \quad (1)$$

čia $f_{t,d}$ – neapdorotas terminų skaičius dokumente, DF – dokumentų, kuriuose yra terminas t, skaičius.

Termino svarbai apskaičiuoti naudojamas IDF (atvirkštinis dokumento dažnis). Atvirkštinis termino dažnis parodo, kokią dalį korpuso dokumentų sudaro terminas. Žodžiai, būdingi tik nedidelei daliai dokumentų įgyja didesnę svarbą nei žodžiai, būdingi visiems dokumentams. Kai kurie angliški terminai, pavyzdžiui, „is“, „an“, „and“ ir kt., pasitaiko dažnai, bet nėra svarbūs [59]. Atvirkštinis termino dažnis (IDF) apskaičiuojamas pagal (2) formulę:

$$idf(t, D) = \log \frac{N}{DF}; \quad (2)$$

čia N – dokumentų skaičius korpuso N = |D|, DF – dokumentų, kuriuose yra terminas t, skaičius.

Turint terminų dažnį (TF) ir atvirkštinį dokumento dažnį (IDF), galime skaičiuoti TF-IDF pagal (3) formulę:

$$tfidf(t, d, D) = tf(t, d) * idf(t, D). \quad (3)$$

Tarkime, kad dokumente yra 200 žodžių, o iš šių 200 žodžių žodis „emisija“ pasikartoja 10 kartų. Termino dažnis bus $10/250 = 0,04$. Tarkime, kad yra 50000 dokumentų ir iš jų tik 500 dokumentų yra žodis „emisija“. Tada,

IDF (emisija) = $50000/500 = 100$, o

TF-IDF (emisija) bus $0,04*100 = 4$

N-grama formuos teksto požymius, skirtus prižiūrimiesiems mašininio mokymosi algoritams. Tai yra n ženklų seka iš pateikto teksto. N reikšmė gali būti 1, 2, 3 ir t.t. Jei laikome, kad n reikšmė yra 1, ji vadinama unigrama, n=2 – bigrama, n=3 – trigramu ir t. t. [59].

Sentimentų klasifikavimas (angl. *Text Classification*). Kiekvieno teksto nuotaikai klasifikuoti naudojami mašininio mokymosi algoritmai arba iš anksto parengti modeliai. Tai galima pasiekti naudojant prižiūrimą mokymąsi, kai modeliai mokomi naudojant pažymėtus duomenis, arba iš anksto paruoštus modelius, kurie išmoko jausmų modelius iš didelių duomenų rinkinių.

Tolesnis apdorojimas. Nuotaikų analizės rezultatai gali būti papildomai apdorojami, pvz., sumuojami nuotaikų balai arba taikomos slenkstinės taisyklės, siekiant klasifikuoti nuotaikas kaip teigiamas, neigiamas arba neutralias.

Įvertinimas. Sentimentų analizės modelio našumas vertinamas naudojant vertinimo metriką, pvz., bendrasis tikslumas (angl. *Accuracy*), tikslumas (angl. *Precision*), jautrumas (angl. *Sensitivity*), F1 balas (angl. *F1 score*).

2.3. Duomenų balansavimo metodai

Klasifikavimo problemos gana dažnas reiškinys dirbant su mašininio mokymosi algoritmais. Dauguma mašininio mokymo algoritmų nelabai gerai veikia su nesubalansuotais duomenų rinkiniais. Nesubalansuoti duomenys – tai tokie duomenų rinkiniai, kuriuose tikslinės klasės stebėjimų pasiskirstymas yra netolygus, t. y. vienai klasei priskirta labai daug stebėjimų, o kitai – labai mažai stebėjimų [82]. Kai skirtumas tarp klasių nežymus, galutiniam rezultatui tai neturi įtakos. Tačiau, kai kiekis tarp klasių žymiai skiriasi, galime gauti prastesnį modelio tikslumą. Šiai problemai spręsti yra taikomos klasių balansavimo technikos, kurias savo straipsniuose nagrinėjo mokslininkai F. Pecorelli et.al. [83], Z. Wang’as et al. [81]:

- Imties didinimas (angl. *Random Oversampling – ROS*). Taikant šią techniką mažumos klasės duomenų kiekis yra padidinamas atsitiktiniu būdu, t.y. dubliuojant mažumos klasės reikšmes, kol bus gautas vienodas įrašų kiekis abiejose klasėse [83].
- Imties mažinimas (angl. *Random Undersampling – RUS*). Metodus panašus į imties didinimo metodą, tik šiuo atveju daugumos imtis yra sumažinama, panaikinant įrašų, kol jis taps vienodas abiejose klasėse [83].
- Sintetinis mažumos imties didinimo metodas – SMOTE (angl. *Synthetic Minority Oversampling Technique*). SMOTE atsitiktinai pasirenka elementą iš mažumos klasės ir nustato jo artimiausius kaimynus [81]. SMOTE metodu nauji atvejai sintetunami iš esamų duomenų. Šis metodas ieško mažumos klasės egzempliorių ir naudoja k artimiausių kaimyną [82]. Atsitiktinai parenkamas kaimynas ir tarp dviejų taškų brėžiama tiesė. Naujas stebėjimas generuojamas parinkus vieną tašką ant tiesės (žr. (4) formulę):

$$s_i = x_i + w * (x_u - x_i); \quad (4)$$

čia s_i – naujas stebėjimas, x_i – imties stebėjimas, x_u – atsitiktinai parinktas stebėjimas iš k kaimynų, w – koeficientas, kai $w \in [0;1]$.

2.4. Leksikonu grįsti metodai

Leksikonai – tai žymių rinkinys, kuriame kiekvienai žymai priskiriamas iš anksto nustatytas balas, nurodantis neutralų, teigiamą ir neigiamą teksto pobūdį [60]. Sentimentų leksikonai – tai žodžių (ir frazių) sąrašai, kurie iš anksto siejami su teigiamomis ir neigiamomis nuotaikomis. Balas, priskiriamas žodžiams ar žodžių junginiui, remiasi poliškumu ir gali įgyti reikšmes nuo -1 iki +1.

Leksikonu paremtas metodas yra labai tinkamas sentimentų analizei atlikti sakinių lygmeniu. Šiam metodui panaudoti nereikia mokymo duomenų, jį galima vadinti nekontroliuojamuoju metodu [61]. Kita vertus, pagrindinis šio metodo trūkumas yra priklausomybė nuo srities, nes žodžiai gali turėti kelias reikšmes ir prasmes, todėl teigiamas žodis vienoje srityje gali būti neigiamas kitoje [62]. Parenkant poliškumą, labai svarbu atsižvelgti į frazės kontekstą.

Žodynu pagrįstą metodą sudaro rankiniu būdu surinktas iš anksto nustatytų nuomonės žodžių sąrašas. Pagrindinė šio metodo prielaida yra ta, kad sinonimai turi tą patį poliškumą kaip ir pagrindinis žodis, o antonimai – priešingą poliškumą [49].

B. Liu teigimu [65] žodynu grįsti metodai gali turėti trūkumų, kai kalbama apie specifines panaudojimo sritis. Taip pat žodžiai apibrėžiami kaip teigiami ar neigiami, nors priklausomai nuo konteksto tas pats žodis gali būti teigiamas ir neigiamas. Kaip pavyzdžiui, „garsiakalbis yra tylus“ – neigiamas, o „automobilis yra tylus“ – teigiamas. Šią problemą gali padėti spręsti korpusu pagrįstas metodas.

Taikant korpusu grindžiamą metodą, remiamasi sintaksės arba bendro pasikartojimo modeliais, taip pat pradinio nuomonės žodžių sąrašu, kad dideliame korpuse būtų galima rasti kitus nuomonės žodžius. Šis metodas pradedamas nuo pradinio nuomonės būdvardžių žodžių sąrašo ir, naudojant juos bei lingvistinių apribojimų arba jungtukų konvencijų rinkinį, nustatomi papildomi nuomonės būdvardžių žodžiai ir jų orientacijos. Vienas iš apribojimų yra susijęs su jungtuku ir (angl. *AND*), kuris sako, kad sujungti būdvardžiai paprastai turi tą pačią orientaciją. Pavyzdžiui, sakinyje „Šis automobilis yra gražus ir erdvus“, jei žinoma, kad „grazus“ yra teigiamas, galima daryti išvadą, kad „erdvus“ taip pat yra teigiamas. Tuo tarpu jungtukai „bet“, „tačiau“, duoda priešingą reikšmę [65].

SenticNet programa yra laisvai prieinamas įrankis sentimentų analizei, kuriuo galima naudotis per nepriklausomą XML repozitoriją arba naudojant aplikacijų programavimo sąsają (API). SenticNet sukurtas taikant grafų gavybos ir daugiamatės skalės metodus. Senticnet 1 (2010 m) susiejo poliariškumo balus su maždaug 6000 sąvokų, o naujausia versija SenticNet 8 (2024 m) naudoja virš 300.000 sąvokų bei pasitelkia dirbtinio intelekto modelius [55]. E. Cambria et al. [56] savo tyrime SenticNet lygino su panašiais įrankiais: rBERT, W2V (angl. *Word2vec*), BoW (angl. *Bag of Words*) bei CGPT (angl. *ChatGPT*). Tyrimas buvo atliekamas sprendžiant tris skirtingus NLP uždavinius (Poliariškumo – nuotaikų analizė; OCEAN – asmenybės prognozavimas; Savižudybės – savižudybių aptikimas). Dauguma atvejų SenticNet gavo geriausią bendrojo tikslumo vertinimo matą (angl. *Accuracy*) lyginint su kitais metodais.

Wordstat Sentimentų žodyną sudaro ~4800 teigiamų ir ~9000 neigiamų žodžių.

Emotikonų žodynas susideda iš 477 emotikonų, pažymėtų priskirtų teigiamai, neigiamai ar neutraliai nuotaikai [52].

Leksikonu grįstų metodų privalumai [48, 49]:

- paprasta suprasti ir įgyvendinti;
- nereikalauja mokymo duomenų ir naudinga tuo atveju, kai mokymo duomenų trūksta;
- veiksmingas tekstams su aiškiai išreikštomis nuotaikomis.

Trūkumai [48, 49]:

- gali netiksliai atspindėti nuo konteksto priklausančias nuotaikas. Pavyzdžiui, sarkazmą, idiomą, neiginius, žargoną, trumpinius gali būti sudėtinga interpretuoti;
- apribota sentimentų leksikono aprėptis ir kokybė;
- sunkiai susidoroja su konkrečiai sričiai būdingomis nuotaikų išraiškomis, nebent leksikonas yra specialiai pritaikytas sričiai.

2.5. Taisyklėmis grįsti metodai

VADER (angl. *Valence Aware Dictionary and sEntiment Reasoner*) – leksikonu ir taisyklėmis grįstas sentimentų analizės metodas, kuriuo tinka analizuoti socialinės medijos žinutes [58]. VADER geba analizuoti tekstą pagal iš anksto mokytą nuotaikų (sentimentų) analizės modelį. O taip pat atsižvelgia į nuotaikos intensyvumą, kurį galima nustatyti pagal didžiosios raidės ir skyrybos ženklus. Pavyzdžiui, visos didžiosios raidės arba šauktukai gali reikšti stipresnę nuotaiką. [62]. Tyrimė analizuojamos žinutės VADER metodu.

VADER yra mažiau išteklių reikalaujantis nuotaikų analizės modelis, kuris naudoja taisyklių rinkinį matematiniam modeliui nurodyti, jo aiškiai nekoduojanč. Palyginti su mašininio mokymosi modeliais, VADER sunaudoja mažiau išteklių, nes nereikia didelių mokymo duomenų kiekių. Išteklius taupantis VADER metodas padeda iššifruoti ir kiekybiškai įvertinti emocijas, esančias srautinėje medijoje, pavyzdžiui, tekste, garso ar vaizdo įrašė [72].

Nagrinėjant F1 balus (klasifikavimo tikslumą) VADER gali paiekti labai aukštų rezultatų (0,96) teisingai priskiria Tviterio žinutes teigiamoms, neutralioms arba neigiamoms klasėms. Taip yra todėl, kad VADER jautriai reaguoja į emocijų poliškumą (ar nuotaika yra teigiama, ar neigiama) ir intensyvumą (kiek teigiama ar neigiama yra nuotaika).

Nagrinėjant tekstą VADER metodu įvertinamas valentiškumo balas. Tai yra balas, skiriamas nagrinėjamam žodžiui remiantis stebėjimais ir patirtimi, o ne gryna logika. Žodžiai „baisus“, „beviltiškas“, „nelaimingas“ – suvokiami kaip neigiamą nuotaiką turintys žodžiai. Kita vertus, tokie žodžiai kaip „nuostabus“, „vertas“, „tinkamas“ reiškia teigiamą nuotaiką. VADER valentingumo balas matuojamas pagal skalę nuo -4 iki +4, kur -4 reiškia didžiausią „neigiamą“ nuotaiką, o +4 – didžiausią „teigiamą“ nuotaiką. Intuityviai galima numanyti, kad 0 reiškia „neutralią“ nuotaiką [63, 72]. Valentingumo rodiklio pavyzdžiai:

- teigiamas: „gerai“ 0,9, „puikiai“ 3,1;
- neigiamas: „siaubingas“ -2,5, emotikonas 😞 -2,2.

Valentingumo balas apskaičiuojamas remiantis penkiomis euristinėmis taisyklėmis, kad įtraukti kiekvienos potekstės įtaką suvokiamam sakinio teksto nuotaikos intensyvumui. Šios euristikos viršija tai, kas paprastai būtų įtraukta į tipinį žodžių maišo modelį. Jos apima žodžių eiliškumui jautresius ryšius tarp terminų.

Žemiau pateikiamos penkios VADER euristicos.

1. **Skyrybos ženklai.** Padidina intensyvumą, sustiprina nuotaiką, nekeisdamas semantinės orientacijos. Pvz. „Lauke karšta!!!“ skamba intensyviau nei „Lauke karšta“.
2. **Didžiųjų raidžių rašymas,** kai norima pabrėžti su jausmais susijusį žodį. Pvz. „Lauke KARŠTA“ perteikia didesnę intensyvumą nei „Lauke karšta“.
3. **Laipsnio modifikatoriai,** dar vadinami intensifikatoriais, stiprinamaisiais žodžiais arba laipsnio prieveiksmiais, (angl. *Booster words*) daro įtaką nuotaikos intensyvumui didindami arba mažindami intensyvumą. Pvz: „Oras yra itin karštas“ yra intensyvesnis nei „Oras yra karštas“, o „Oras yra šiek tiek karštas“ sumažina intensyvumą.
4. **Kontrastinis jungtukas** „bet“ signalizuoja apie sentimentų poliariskumo pasikeitimą, kai dominuoja po jungtuko einančio teksto sentimentai. Pvz: „Oras karštas, bet jis pakenčiamas“ turi mišrų sentimentą, o antroji pusė diktuoja bendrą vertinimą.
5. **Poliškumo neigimas.** Neigimas pakeičia teksto poliariskumą. Pavyzdžiui, paneigtas sakinytis būtų toks: „Oras tikrai nėra toks karštas“.

VADER sudėtinis balas (angl. *Compound score*) nuotaikoms analizuoti apskaičiuojamas sumuojant kiekvieno leksikono žodžio valentingumo balus, pakoreguotus pagal taisykles ir normalizuotus taip, kad būtų nuo -1 (labiausiai neigiamas) iki +1 (labiausiai teigiamas). Tai naudingiausia metrika, kai norima gauti vieną vienmatį tam tikro sakinio nuotaikos matą.

Duomenų normalizavimas atliekamas pagal matą (žr. (5) formulę):

$$x = \frac{x}{\sqrt{x^2 + \alpha}}; \quad (5)$$

čia x – sudedamųjų žodžių valentingumo balų suma, α – normalizavimo konstanta (numatytoji reikšmė yra 15)

VADER sentimentų analizė gali būti atliekama *Python* kalba panaudojant *Jupyter* notebook aplinką. Instaliuojami paketai *vaderSentiment*, *nltk*. Darbo tyrimo dalyje bus detalčiau nagrinėjamas šis metodas.

Panaudojus šias bibliotekas apskaičiuojamas poliškumo matas (angl. *Polarity score*), kuriame nurodoma neigiama, teigiama, neutrali reikšmė, bei įvertinams sudėtinis balas (angl. *Compound*). Sentimentai vertinami pagal apskaičiuotą sudėtinį balą:

- teigiama nuotaika 😊, jei šio balo reikšmė $\geq 0,05$;
- neigiama 😞, jei $\geq -0,05$;
- neutrali 😐, jei reikšmė daugiau už -0,05 ir mažiau už 0,05

2.6. Mašininio mokymosi metodai

Mašininis mokymasis – dirbtinio intelekto metodų klasė, kuriai būdingas ne problemos sprendimas, o mokymasis [57]. Iš 13 pav. pateikto klasifikavimo medžio, [49, 57] matyti, kad galimi prižiūrėti (angl. *Supervised*) ir neprižiūrėti (angl. *Unsupervised*) mašininio mokymo metodai.

Taikant prižiūrėtus metodus naudojamas didelis skaičius pažymėtų mokymo dokumentų [57]. Prižiūrėto mokymosi algoritmai sukuria matematinio duomenų rinkinio modelį, kuriame yra įvesties duomenys (x) ir pageidaujami išvesties duomenys (Y).

Neprižiūrimo mokymosi algoritmai yra tokie, kurie turi įvesties duomenis (x), bet neturi jiems priskirtų išvesties reikšmių. Šie algoritmai padeda rasti struktūras duomenyse, kurie nebuvo pažymėti, suklasifikuoti ar suskirstyti į kategorijas [73]. Klasterizavimo (angl. *Clustering*), asociacijų (angl. *Association rules*), dimensijų mažinimo uždaviniai gali būti priskiriami neprižiūrimo mokymosi metodams.

Sentimentų analizei tirti naudojami gerai žinomi mašininio mokymosi algoritmai:

- sprendimų medžio klasifikatorius;
- linijiniai klasifikatoriai;
- taisyklėmis pagrįsti klasifikatoriai;
- tikimybiniai klasifikatoriai.

Populiariausi iš tikimybinių klasifikatorių yra

- Naivusis Bajesas klasifikatorius (angl. *Naive Bayes*, toliau NB);
- Bajeso tinklo (angl. *Bayesian Network*, BN);
- maksimalios entropijos klasifikatorius (ME).

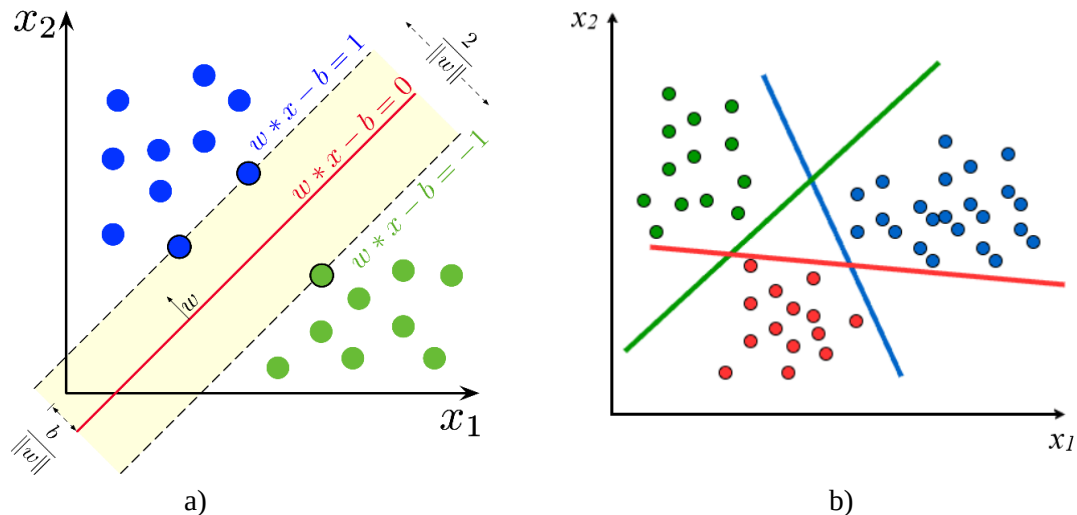
Tyrime nagrinėjamas atraminių vektorių metodas (SVM) ir Naivaus Bajeso klasifikavimo metodas (NB).

2.6.1. Atraminių vektorių klasifikatorius (SVM)

Atraminių vektorių klasifikatorius (angl. *Support Vector Machine*, toliau SVM) – tinkamas metodas sentimentų analizei tirti. Prižiūrimas mašininio mokymosi algoritmas SVM, tiria duomenis ir nustato šablonus, naudojamus skirstant į kategorijas. SVM turi galimybę identifikuoti atskiras hiperplokštumas, kurios padeda maksimaliai padidinti skirtumą tarp įvairių klasių [45].

Klasifikavimo naudojant SVM algoritmą metodika pradedama transformuojant tekstinius duomenis į svorius. Šie svoriai dažnai sujungiami, kad susidarytų TF-IDF reikšmės, tiesiog padauginus jas kartu.

Pasak M. D. Devika – tai netikimybinis klasifikatorius, kuriame reikalingas didelis treniravimo duomenų rinkinys. Tai atliekama klasifikuojant taškus ir atskiriant juos hiperplokštuma [45] ir atvaizduojama 17 pav.



17 pav. Atraminų vektorių klasifikatoriaus: a) 2 klasių klasifikavimo pavyzdys, b) daugiau kaip dviejų klasių klasifikavimo pavyzdys [45]

Duomenų taškai. Čia yra pavaizduoti duomenų taškai dvimatėje erdvėje (17 pav. a), kur x_1 ir x_2 yra savybės. Mėlyni objektai atstovauja vienai klasei, o žali – kitai.

Hiperplokštuma (17 pav. a) – raudona linija): Tai sprendimų riba, kuri atskiria du skirtingus klasių taškus. Idealus atvejis yra, kai ši linija yra taip padėta, kad maksimaliai padidintų atstumą tarp artimiausių skirtingų klasių taškų. SVM naudoja lengvos paraštės, griežtos paraštės ar branduolio metodus. Bendru atveju sprendimo funkciją aprašo (6) formulė:

$$D_i(x) = w_i^T \phi(x) - b_i; \quad (6)$$

čia w_i – svorių vektorius, $\phi(x)$ – funkcija, kuri braižo x į dimensijos požymių erdvę, o b_i – poslinkio (angl. *bias*) parametras

Paraštės – tai dvi lygiagrečios linijos hiperplokštumos abiejose pusėse. Jos apibrėžia erdvę tarp klasių ir yra tolygios nuo hiperplokštumos. Šios linijos matematiškai aprašomos (7) arba (8) formule

$$w_i^T \phi(x) - b_i = 1 \quad (7)$$

arba

$$w_i^T \phi(x) - b_i = -1 \quad (8)$$

Atraminiai vektoriai (angl. *support vectors*) – tai duomenų taškai, esantys arčiausiai sprendimo ribos. SVM modelis juos naudoja nustatydamas optimaliausią sprendimo ribą. Atraminiai vektoriai yra būtini paraščių apskaičiavimui ir modelio atsparumui duomenų pokyčiams.

Vektorius w ir atstumas. Vektorius w nurodo hiperplokštumos orientaciją. Jei hiperplokštuma $D_i(x) = 0$, turime optimaliai skiriančią hiperplokštumą. Jei klasifikavimo uždavinyje atskiriame mokymo duomenis, duomenys priklauso klasei i , tenkinančiai $D_i(x) \geq 1$. Kitos klasės duomenys tenkina sąlygą $D_i(x) \leq -1$. Paraštei maksimizuoti, Euklido norma, t. y. $\frac{1}{2} ||w||^2$ turi artėti į minimumą.

SVM modelio tikslas yra rasti tokį svorių vektorių w (žr. (9) formulę) ir poslinkio parametą b (žr. (10) formulę), kad būtų maksimaliai padidintas atstumas tarp atraminių vektorių ir hiperplokštumos, taip užtikrinant geriausią klasifikacijos sprendimą.

$$w = \sum_{i=1}^M \alpha_i y_i x_i \quad (9)$$

ir

$$b = y_i - w^T x_i; \quad (10)$$

čia α_i – neneigiami Lagranžo daugikliai.

Modelis mokomas taip, kad artimiausi skirtingų klasių taškai (atraminiai vektoriai) būtų kiek įmanoma toliau nuo hiperplokštumos. SVM yra labai efektyvus, kai yra aiški atskiriamoji riba tarp klasių.

Šio metodo tyrinėtojai M. Wankhade, A. C. S. Rao ir C. Kulkarni [49] padarė išvadas, apie modelio privalumus:

- SVM – populiariausias sentimentų analizės tyrimo būdas;
- duoda gerą rezultatą dideliems duomenų rinkiniams.

Modelio trūkumus paminėjo šiuos:

- ilgas didelių duomenų rinkinių mokymo laikas;
- tikslus modelio derinimas yra gana sudėtingas ir ilgai trunkantis procesas.

Kai duomenys yra netiesiniai ir jų negalima atskirti tiesia linija, SVM šiai problemai spręsti naudoja „branduolio triuką“. Naudojant branduolio triuką duomenys transformuojami į didesnės dimensijos erdvę naudojant matematinę funkciją, vadinamą branduolio funkcija.

Ši transformacija leidžia SVM surasti kreivą arba netiesinę ribą, kuri gali atskirti duomenis. Naudodamas įvairių tipų branduolio funkcijas, SVM gali apdoroti įvairių rūšių netiesinius duomenų modelius. Šis metodas padeda SVM užfiksuoti sudėtingesnius ryšius ir dėsningumus, todėl jis yra veiksmingas net tada, kai duomenys nėra tiesiškai atskiriami.

Įgyvendinant SVM, reikia atsižvelgti į du esminius parametrus: C ir γ (*gama*). C yra minkštos paraštės optimizavimo parametras. Minkštai paraštei minimizuoti taikoma (11) formulė. Parametru C nustatomas kompromisas tarp tikslaus mokymo taškų klasifikavimo ir tolygios sprendimo ribos pasiekimo. Didesnė C reikšmė skatina modelį pasirinkti daugiau atraminių vektorių, o mažesnė C reikšmė – mažiau atraminių vektorių.

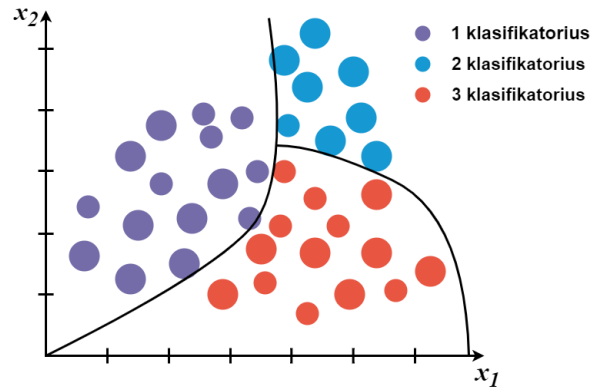
$$Q(w, b, \xi) = \frac{1}{2} ||w||^2 + \frac{c}{p} \sum_{i=1}^M \xi_i^p; \quad (11)$$

čia p – bauda (angl. *penalty*), o ξ_i – neneigiami laisvieji kintamieji.

Darbo tiriamojoje dalyje išbandytas modelis su skirtingomis C parametro reikšmėmis, siekiant gauti optimalius rodiklius.

2.6.2. Naivusis Bajesas klasifikatorius (NB)

Naivaus Bajeso (NB) klasifikatorius pats paprasčiausias ir dažniausiai naudojamas klasifikatorius, kuris suskirsto naujus pavyzdžius į vieną iš anksčiau nustatytų klasių (žr. 18 pav.). Naivus Bajesas yra tikimybinis klasifikatorius, pagrįstas Bajeso teorema, kai požymiai yra nepriklausomi vienas nuo kito.



18 pav. Klasifikavimas Naivaus Bajeso metodu [80]

Pagal Bajeso teoremą nustatoma įvykio tikimybė, atsižvelgiant į kito jau įvykusio įvykio tikimybę. Bajeso teorema matematiškai išreikšta tokia lygtimi (žr. (12) formulę) [80]:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)}; \quad (12)$$

čia A ir B – įvykiai, o $P(B) \neq 0$ ir

$P(B)$ – **įrodymai** (angl. *Evidence*), kurie sudaro išankstinę mūsų įvykio sąlygą. Bandome rasti įvykio A tikimybę, jei įvykis B yra teisingas. Įvykis B vadinamas įrodymu;

$P(A_i)$ yra **išankstinė tikimybė** (angl. *Priori*), t. y. įvykio tikimybė prieš pamatant įrodymą. Įrodymas yra nežinomo atvejo (šiuo atveju tai yra įvykis B) atributo reikšmė;

$P(A_i|B)$ – B **posteriori tikimybė**, t. y. sąlyginė įvykio A tikimybė po to, kai matomi įrodymai;

$P(B|A_i)$ yra **tikimybės tikimybė** (angl. *Likelihood*), t. y. tikimybė, kad hipotezė pasitvirtins remiantis įrodymais.

Pagal tai (12) formulę gali būti perfrazuojama į (13) formulę:

$$\text{posteriori tikimybė} = \frac{\text{išankstinė tikimybė} * \text{tikimybės tikimybė}}{\text{įrodymas}}. \quad (13)$$

Daroma prielaida, kad kiekvienas įvesties požymis nepriklausomai ir vienodai prisideda prie galutinio rezultato.

Duotam nepriklausomam n dydžio požymių vektoriui X su požymiais $x_1, x_2, x_3, \dots, x_n$ prognozuojame naujo atvejo klasę Y_{pred} (žr. (14) formulę):

$$Y_{pred} = \operatorname{argmax}_{y \in Y} P(y|x_1, x_2, \dots, x_n); \quad (14)$$

čia Y – galimų klasių aibė.

Pritaikius Bajeso teoremą, gauname (žr. (15) formulę):

$$Y_{pred} = \operatorname{argmax}_{y \in Y} \frac{P(x_1, x_2, \dots, x_n | y) P(y)}{P(x_1, x_2, \dots, x_n)}; \quad (15)$$

Kadangi $P(x_1, x_2, \dots, x_n)$ yra pastovus, Y_{pred} galima užrašyti išraiška (žr. (16) formulę):

$$Y_{pred} = \operatorname{argmax}_{y \in Y} P(x_1, x_2, \dots, x_n | y) P(y). \quad (16)$$

Norint apskaičiuoti $P(x_1, x_2, \dots, x_n | y)$, reikia rasti duomenis, kurie priklauso y klasei su tikslia $x_1, x_2, \dots, x_n$ kombinacija. Taigi, darome prielaidą, kad $x_1, x_2, \dots, x_n$ yra nepriklausomi požymių įvesties duomenys. Atsižvelgiant į klasę y , taikome sąlyginį nepriklausomumą (žr. (17) formulę):

$$P(x_1, x_2, \dots, x_n | y) = P(x_1 | y) P(x_2 | y) \dots P(x_n | y) = \prod_i P(x_i | c); \quad (17)$$

todėl papildytas Naivaus Bajeso klasifikatorius užrašomas (18) formule:

$$Y_{pred} = \operatorname{argmax}_{y \in Y} P(y) \prod_i P(x_i | y). \quad (18)$$

Šio metodo privalumai: veikia greitai, tinka kelių klasių numatymo problemoms spręsti. Jei pasitvirtina prielaida dėl funkcijų nepriklausomumo, gali veikti geriau nei kiti modeliai ir reikalauja mažiau mokymo duomenų. Gerai tinka kategoriniams kintamiesiems.

Trūkumai: algoritmas daro prielaidą, kad visos savybės yra nepriklausomos, o tai realiame gyvenime pasitaiko retai. Algoritmą reikia išglodinti (angl. *smoothing*), kad būtų galima išspręsti problemas, kylančias dėl testavimo duomenų funkcijų, kurių nebuvo mokymo duomenyse. Algoritmas gali pateikti klaidingus įvertinimus, pvz. Naivus Bajesas daro prielaidą, kad bruožai nepriklausomai prisideda prie klasės atsiradimo tikimybės. Kai taikome šį metodą duomenų rinkiniui, apimančiame tarpusavyje susijusias funkcijas, vis tiek reikia vienodai prisidėti prie visų savybių, todėl klaidingai įvertinama tikimybė [80].

2.7. Tikslumo vertinimo metrikos

Bendrasis tikslumas (angl. *Accuracy*), **tikslumas** (angl. *Precision*), **jautrumas** (angl. *Sensitivity*), **F1 balas** (angl. *F1 score*) yra vertinimo metrikos, naudojamos mašininio mokymosi modelių veikimui vertinti, ypač atliekant klasifikavimo užduotis.

Šiose metrikose dominuoja kriterijai, kurie padeda nustatyti klasifikuotų žinučių tikslumą:

- TP (angl. *True Positive*) – teigiamų žinučių skaičius, kurios buvo teisingai klasifikuotos;
- TN (angl. *True Negative*) – neigiamų žinučių kiekis, kuris buvo teisingai klasifikuotas;
- FP (angl. *False Positive*) – neteisingai klasifikuotų teigiamų žinučių;
- FN (angl. *False Negative*) – neteisingai klasifikuotų neigiamų žinučių.

Šiuos kriterijus atvaizduojame sumaišymo matricoje, norint nustatyti kiek duomenų pagal klases modelis prognozuoja teisingai (žr. 19 pav.) Sumaišymo matrica – tai lentelė, kuri dažnai naudojama vertinant klasifikavimo modelio veiksmingumą, kai yra žinomos mokomųjų testo duomenų reikšmės [49].

		Tikrosios reikšmės	
		Teigiami	Neigiami
Spėjamos reikšmės	Teigiami	TP	FP
	Neigiami	FN	TN

19 pav. Sumaišymo matricos pavyzdys 2 klasėms

Sumaišymo matricą galima naudoti sprendžiant daugiau nei 2 klasių klasifikavimo uždavinius. Tokiu atveju matrica turės daugiau nei dvi eilutes ir stulpelius. Kiekviena eilutė reiškia spėjamas reikšmes, o stulpelis – faktines reikšmes [79] (žr. 20 pav.)

		Tikrosios reikšmės		
		Teigiami (1)	Neutralūs (0)	Neigiami (-1)
Spėjamos reikšmės	Teigiami (1)	700	300	0
	Neutralūs (0)	200	8300	100
	Neigiami (-1)	0	100	300

20 pav. Sumaišymo matricos pavyzdys 3 klasėms

Bendrasis tikslumas (angl. *Accuracy*). Šiuo rodikliu matuojama modelio teisingų prognozių dalis visame duomenų rinkinyje. Jis apskaičiuojamas kaip teisingų teigiamų (TP) ir teisingų neigiamų (TN) rezultatų santykis su bendru mėginimų skaičiumi [75]:

$$\text{Bendrasis tikslumas} = \frac{TP+TN}{TP+TN+FP+FN}; \quad (19)$$

Tikslumas (angl. *Precision*) parodo teisingų teigiamų prognozių dalį tarp visų teigiamų modelio prognozių [75]:

$$\text{Tikslumas}_{\mu} = \frac{TP_1 + \dots + TP_n}{TP_1 + \dots + TP_n + FP_1 + \dots + FP_n}; \quad (20)$$

čia ir toliau n – klasių skaičius.

Jautrumas (angl. *Sensitivity*) – tikrojo teigiamo rezultato rodiklis, parodo tikrojo teigiamo rezultato prognozių dalį tarp visų faktinių teigiamų atvejų. Jis apskaičiuojamas kaip TP sumos ir TP bei klaidingų neigiamų rezultatų (FN) sumos santykis:

$$\text{Jautrumas}_{\mu} = \frac{TP_1 + \dots + TP_n}{TP_1 + \dots + TP_n + FN_1 + \dots + FN_n}; \quad (21)$$

F1 balas (angl. *F1 score*) yra tikslumo (*Precision*) ir jautrumo (*Sensitivity*) pusiausvyrą užtikrinantis rodiklis. Jis apskaičiuojamas kaip tikslumo ir atšaukimo harmoninis vidurkis. F1 balas naudingas, kai siekiama pusiausvyros tarp didelio tikslumo ir didelio jautrumo, nes baudžiama už itin neigiamas bet kurio komponento reikšmes.

$$F1_{\mu} = 2 * \frac{TP_1 + \dots + TP_n}{TP_1 + \dots + TP_n + FP_1 + \dots + FP_n + FN_1 + \dots + FN_n}; \quad (22)$$

Bendrasis tikslumas vertina bendrą modelio prognozių teisingumą, o tikslumas ir jautrumas – atitinkamai teigiamų ir neigiamų prognozių kokybę. F1 balas užtikrina pusiausvyrą tarp tikslumo ir jautrumo, todėl yra išsamesnė klasifikavimo modelių vertinimo metrika [75].

3. Tyrimo eiga ir rezultatai

Tyrimas yra paremtas matematiniais skaičiavimais, tačiau taip pat orientuotas į verslo įmones ir turintis praktinį panaudojimą. *Python* kalba panaudojant *Jupyter notebook* aplinką atliktas duomenų apdorojimas ir matematiniai skaičiavimai. Atlikus duomenų apdorojimą, duomenys saugomi *.parquet formatu. Po to *Microsoft Power BI* įrankiu sukurta vartotojo sąsaja ir duomenų vizualizavimas. *Power BI* – lengvai valdoma, interaktyvi vartotojo sąsaja, vizualiai pateikianti gautus rezultatus švieslentėje.

3.1. Duomenų gavybos procesas

Tyrimui panaudotos 778 žinutės, iš kurių *E* – 432 žinutės, *S* – 207, *G* – 139. Žinučių publikavimo laikotarpis 2015-2024 m.

Žinutės surinktos *Twitter* platformoje, paminint mados sektoriaus įmones ir panaudojant ESG temoje nagrinėtus raktažodžius. *Twitter* paieškos lauke įvedama užklausa, kurios bendrinė schema atrodo taip:

Brand1 OR Brand2 #ESG OR #HumanRights OR #TextileWaste OR #PlasticBags

Tyrimo nagrinėjamos mados industrijos kompanijos iš TOP 100 sąrašo [71]. Į šį sąrašą patenka tiek greitosios mados, tiek ir prabangos prekiniai ženklai. Galima tikėtis skirtingų atsiliepimų apie abiejų kategorijų kompanijas, dėl to žinutės užklausa apima tiek vienos, tiek kitos kategorijos atsiliepimus. Analizuojama kiekviena ESG tema atskirai, t.y. pateikiama užklausa su *E* temos raktažodžiais, *S* ir *G*.

Dėl *Twitter* duomenų apribojimų politikos, duomenys buvo renkami rankiniu būdu pateikiant užklausas. Gautas žinutės užsaugomas tekstiniu pavidalu *.txt formatu.

Užklausa aplinkosaugos tema (angl. *E* – *Environment*):

@GAP OR @Adidas OR @Nike OR @Lululemon #Carbon OR #Emission OR #Renewable OR #Waste OR #Water OR #Plasticbags OR #Manufacture OR #recycle OR #sustainable OR #Climate OR #Supplychain OR #reuse OR #circulareconomy OR #depletion

Atlikus vieną paiešką aplinkosaugos tema, keičiami kompanijų pavadinimai ir taip pildomas duomenų rinkinys. Pateikus užklausą *Twitter* platformoje gaunamas rezultatas pateiktas 21 pav.



21 pav. Twitter žinutės pavyzdys aplinkosaugos tema

Taigi matome, kad žinutėje šalia @ simbolio figūruoja kompanijos pavadinimas, o šalia # – bent vienas iš užklausoje paminėtų raktažodžių. Analizuojant žinutes surenkame ne tik paminėtą # raktinį žodį, bet visus žinutėje paminėtus raktinius žodžius.

Tekstiniame dokumente žinutės atvaizduojamos tokiu formatu (žr. 22 pav.)

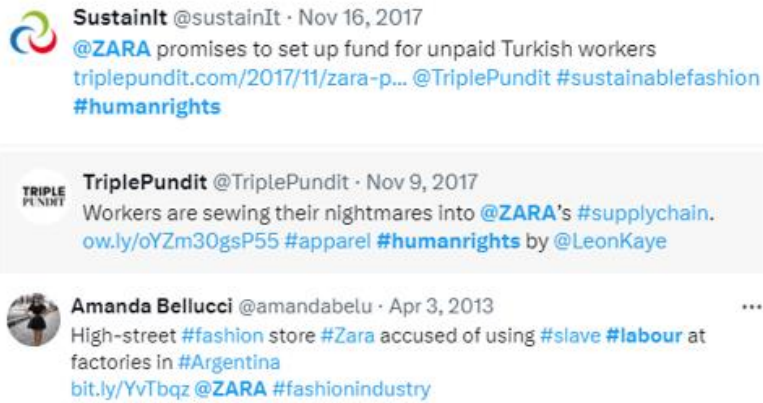
<i>Fur Free Retailer</i> <i>@FurFreeRetailer</i> <i>.</i> <i>Feb 17, 2021</i> <i>We are thrilled to welcome sportswear icon</i> <i>@adidas</i> <i>as the 1500th brand to join our program!</i> <i>With today's commitment, the sports apparel giant joins a new wave of major brands announcing #fur bans and championing #sustainable fashion 🐾 https://tinyurl.com/16yep2y6 #adidasfurfree</i>

22 pav. Twitter žinutės pavyzdys paruoštas teksto apdorojimui

Analogiškai atlikta paieška socialinės atsakomybės (angl. *S – Social*) tema, pakeičiant raktažodžius į šiuos:

@Nike OR @Zara OR @Adidas #Workforce OR #Humancapital OR #Labour OR #Privacy OR #Recruitment OR #Accident OR #Safety OR #Injury OR # HumanRights OR #Wages

Surandamos žinutės, kurios kopijuojamos į tekstinio tipo dokumentą *.txt. Keletas tokių žinučių pavyzdžių pateiktos 23 pav.



23 pav. Twitter žinučių pavyzdžiai socialinės atsakomybės tema

Žinučių paieška valdymo srityje (angl. *G – Governance*) atlikta su raktažodžiais:

@GAP OR @Adidas OR @Nike OR @Lululemon #Governance OR #Corporate OR #Stockmarket OR #Supplychain OR #Shareholder OR #Corruption OR #Transparency OR #BusinessEthics OR #Investment OR #Finance

Žinučių pavyzdžiai valdymo tema pateikti 24 pav.



24 pav. Twitter žinučių pavyzdžiai valdymo tema

3.2. Teksto apdorojimas ir požymių išskyrimas

Pirmame duomenų apdorojimo etape, žinučių tekstas suskirstomas į duomenų struktūras (angl. *Dataframe*), tokias kaip žinutės skelbėjas, paskelbimo data, žinutės tekstas, @kompanija, #raktažodis. Toliau atliekamas žinučių teksto apdorojimas ir požymių ištraukimas.

Šiam procesui panaudotos *Python* bibliotekos ir paketai yra šie:

- biblioteka **pandas** – duomenų rinkinių tvarkymui;
- biblioteka **pprint**, paketas **pprint** – gražiai atspausdinti bet kokias *Python* duomenų struktūras;
- biblioteka **nltk**, paketas **corpus**, procedūra **stopwords** – šalina stop žodžius iš žinučių;
- biblioteka **nltk**, paketas **tokenize**, procedūra **word_tokenize** – žodžių išskyrimas;

- biblioteka **re** – Regular expression operations – šalina specifinius simbolius, URL nuorodas ir pan.;
- biblioteka **unidecode**, paketas **unidecode** – šalina simbolius ir emotikonus iš žinučių;
- biblioteka **pyarrow** – eksportuoti į failo tipą *.parquet;
- biblioteka **collections**, paketas **counter** – skirtas objektų skaičiavimui;
- biblioteka **sklearn**, paketas **preprocessing**, procedūra **normalize** – vektorių normavimui.

Teksto apdorojimo tvarka ir rezultatas pateiktas 4 lentelėje.

4 lentelė. Teksto apdorojimo etapai su atsitiktinai pasirinktomis dviem *Twitter* žinutėmis

Apdorojimo žingsnis	Rezultatas
Nekoreguota žinutė	@Burberry will source 100% #RenewableElectricity this year! 🇬🇧 #RE100 https://edie.net/news/
	@Burberry becomes the first #luxury #brand to commit to being #ClimatePositive by 2040, including actions beyond its value chain 🇬🇧
Iš duomenų rinkinio šalinami žinučių dublikatai	
URL nuorodų šalinimas	@Burberry will source 100% #RenewableElectricity this year! 🇬🇧 #RE100
	@Burberry becomes the first #luxury #brand to commit to being #ClimatePositive by 2040, including actions beyond its value chain 🇬🇧
Skyrybos ženklų, simbolių, skaičių, emotikonų ir pašalinių tarpų šalinimas	Burberry will source RenewableElectricity this year RE
	Burberry becomes the first luxury brand to commit to being ClimatePositive by including actions beyond its value chain
Raidžių vertimas į mažąsias	burberry will source renewableelectricity this year re
	burberry becomes the first luxury brand to commit to being climatepositive by including actions beyond its value chain
Stop žodžių šalinimas	burberry, source, renewableelectricity, year
	burberry, becomes, first, luxury, brand, commit, climatepositive, including, actions, beyond, value, chain
Tokenizavimas	'burberry', 'source', 'renewableelectricity', 'year'
	'burberry', 'becomes', 'first', 'luxury', 'brand', 'commit', 'climatepositive', 'including', 'actions', 'beyond', 'value', 'chain'
Kamienavimas	'burberri', 'sourc', 'renewableelectr', 'year'
	'burberri', 'becom', 'first', 'luxuri', 'brand', 'commit', 'climateposit', 'includ', 'action', 'beyond', 'valu', 'chain'

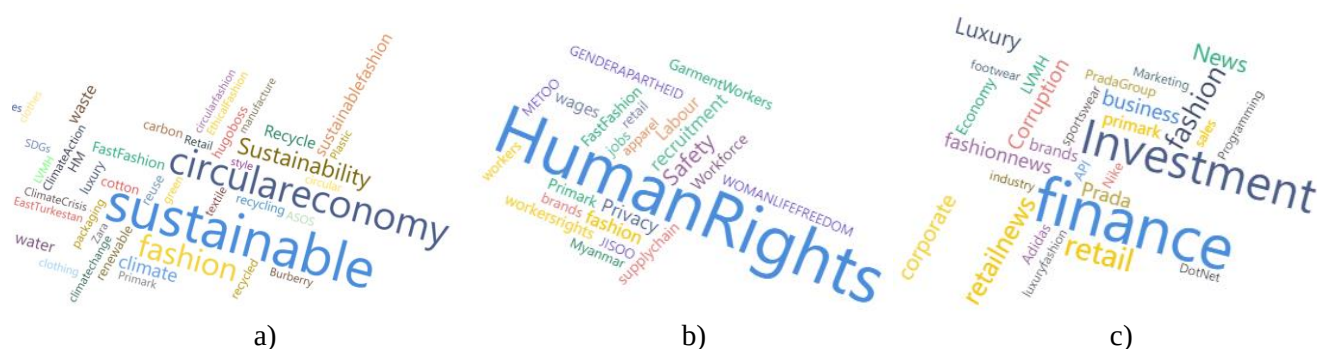
Taip apdorotos žinutės toliau gali būti analizuojamos sentimentų analizės metodais, skaičiuojami žodžių dažniai ir pan.

Žodžių vektorius tyrimui naudojami žodžiai su # simboliu. Tai yra raktažodžiai, kurie atskleidžia pokalbio esminę temą. 25 pav. matome, dažniausiai pasikartojančius žodžius kiekvienai iš sričių.



25 pav. Dažniausiai minimų žodžių stulpelinė pasiskirstymo diagrama: a) aplinkosaugos srityje, b) socialinės atsakomybės srityje, c) valdymo srityje

Žodžių debesis – vienas iš duomenų atvaizdavimo būdų, taip pat parodo dažniausiai panaudotus žodžius, kurie debesyje vaizduojami didžiausi (žr. 26 pav.). Pasirinkti žodžiai, kurie žinutėse pasikartojo daugiau nei 10 kartų.



26 pav. Žodžių debesis, sudarytas iš dažniausiai paminėtų žodžių: a) aplinkosaugos srityje, b) socialinės atsakomybės srityje, c) valdymo srityje

Raktiniai žodžiai, pasikartojantys paieškoje daugiau ar mažiau susiję tarpusavyje. 27 pav., 28 pav., 29 pav. pateiktos atstumų matricos parodo, kaip stipriai susiję žodžiai, t.y. kaip dažnai ieškant vieno žodžio paminimas ir kitas raktažodis toje pačioje žinutėje.

Atstumų matrica analizuoja raktinių žodžių tarpusavio sąryšį. Kuo mažesnė reikšmė, tuo žodžiai artimesni, t.y. atstumas tarp jų mažiausias. Kuo mažiau raktinių žodžių žinutėje, tuo jie artimesni. Gauta reikšmė -1,00 rodo, kad šie žodžiai nepasitaikė kartu nei vienoje žinutėje.

	carbon	emission	renewable	waste	water	plasticbags	manufacture	recycle	sustainable	climate	suplychain	reuse
carbon	0.000000	0.251695	0.168374	0.251695	0.251695	0.251695	0.251695	0.251695	0.168391	0.251695	0.251695	0.251695
emission	0.251695	0.000000	0.251695	0.251695	0.251695	0.251695	0.251695	0.251695	0.251695	0.169135	0.251695	0.251695
renewable	0.168374	0.251695	0.000000	0.251695	0.251695	0.251695	0.251695	0.251695	0.168150	0.168150	0.251695	0.251695
waste	0.251695	0.251695	0.251695	0.000000	0.168011	0.251695	0.251695	0.167895	0.126390	0.251695	0.251695	0.168037
water	0.251695	0.251695	0.251695	0.168011	0.000000	0.251695	0.251695	0.251695	0.251695	0.170269	0.251695	0.251695
plasticbags	0.251695	0.251695	0.251695	0.251695	0.251695	0.000000	0.251695	0.251695	0.251695	0.251695	0.251695	0.251695
manufacture	0.251695	0.251695	0.251695	0.251695	0.251695	0.251695	0.000000	0.251695	0.251695	0.251695	0.251695	0.251695
recycle	0.251695	0.251695	0.251695	0.167895	0.251695	0.251695	0.251695	0.000000	0.101056	0.251695	0.251695	0.072703
sustainable	0.168391	0.251695	0.168150	0.126390	0.251695	0.251695	0.251695	0.101056	0.000000	0.127521	0.251695	0.126180
climate	0.251695	0.169135	0.168150	0.251695	0.170269	0.251695	0.251695	0.251695	0.127521	0.000000	0.251695	0.251695
suplychain	0.251695	0.251695	0.251695	0.251695	0.251695	0.251695	0.251695	0.251695	0.251695	0.251695	0.000000	0.251695
reuse	0.251695	0.251695	0.251695	0.168037	0.251695	0.251695	0.251695	0.072703	0.126180	0.251695	0.251695	0.000000

27 pav. Atstumų matrica su raktiniais žodžiais aplinkosaugos (E) tema

Iš 27 pav. matome, kad žodžių poros *sustainable – carbon*, *circulareconomy – sustainable*, *recycle – sustainable*, *reuse – recycle* dažnu atveju dalyvavo kartu žinutėse kaip susiję raktiniai žodžiai.

Analogiškai tikrinami raktiniai žodžiai kitame duomenų rinkinyje – socialinės atsakomybės (S) tema. Rezultatas pateiktas 28 pav.

	workforce	humancapital	labour	privacy	recruitment	accident	safety	injury	wages	humanrights
workforce	0.000000	0.505882	0.505882	0.505882	0.505882	0.505882	0.505882	0.505882	0.505882	-1.000000
humancapital	0.505882	0.000000	0.505882	0.505882	0.505882	0.505882	0.505882	0.505882	0.505882	-1.000000
labour	0.505882	0.505882	0.000000	0.505882	0.505882	0.505882	0.505882	0.505882	0.255237	-1.000000
privacy	0.505882	0.505882	0.505882	0.000000	0.505882	0.505882	0.505882	0.505882	0.505882	-1.000000
recruitment	0.505882	0.505882	0.505882	0.505882	0.000000	0.505882	0.252395	0.505882	0.505882	-1.000000
accident	0.505882	0.505882	0.505882	0.505882	0.505882	0.000000	0.505882	0.505882	0.505882	-1.000000
safety	0.505882	0.505882	0.505882	0.505882	0.252395	0.505882	0.000000	0.505882	0.505882	0.510638
injury	0.505882	0.505882	0.505882	0.505882	0.505882	0.505882	0.505882	0.000000	0.505882	-1.000000
wages	0.505882	0.505882	0.255237	0.505882	0.505882	0.505882	0.505882	0.505882	0.000000	0.509615
humanrights	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	0.510638	-1.000000	0.509615	0.000000

28 pav. Atstumų matrica su raktiniais žodžiais socialinės atsakomybės (S) tema

Iš 28 pav. matome, kad artimos žodžių poros buvo *wages – labour* ir *safety – recruitment*, o kitos žodžių poros pasitaikė kartu vienodai retai.

	governance	corporate	stockmarkets	supplychain	shareholder	corruption	investment	investing	finance
governance	0.000000	1.916667	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000
corporate	1.916667	0.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000
stockmarkets	-1.000000	-1.000000	0.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	1.900000
supplychain	-1.000000	-1.000000	-1.000000	0.000000	-1.000000	-1.000000	1.933333	-1.000000	1.941176
shareholder	-1.000000	-1.000000	-1.000000	-1.000000	0.000000	-1.000000	-1.000000	-1.000000	-1.000000
corruption	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	0.000000	-1.000000	-1.000000	-1.000000
investment	-1.000000	-1.000000	-1.000000	1.933333	-1.000000	-1.000000	0.000000	1.980583	5.909071
investing	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	1.980583	0.000000	5.854879
finance	-1.000000	-1.000000	1.900000	1.941176	-1.000000	-1.000000	5.909071	5.854879	0.000000

29 pav. Atstumų matrica su raktiniais žodžiais valdymo (G) tema

29 pav. matome, kad tokios žodžių poros kaip governance – corporate, investment – investing, finance – shareholder pasitaikė kartu vienoje žinutėje, o kiti raktiniai paieškos žodžiai tarpusavyje nesudarė porų.

3.3. Sentimentų analizė naudojant Twitter žinutes

3.3.1. VADER sentimentų analizė

Metodologinėje darbo dalyje buvo aptartos penkios VADER euristicos. Panaudojant šias taisykles išanalizuota žinutė, kad įsitikinti kaip keičiasi rezultatas, priklausomai nuo skyrybos ženklų panaudojimo, būdvardžių sustiprinimo, didžiųjų raidžių, poliškuno pakeitimo ir pan. Rezultatai pateikiami 5 lentelėje. Žinutės nuotaika vertinama pagal apskaičiuotą sudėtinį balą – compound, t.y.,

- teigiama nuotaika 😊, jei reikšmė $\geq 0,05$
- neigiama 😞, jei $\geq -0,05$
- neutrali 😐, jei reikšmė daugiau už $-0,05$ ir mažiau už $0,05$

5 lentelė. VADER euristicų panaudojimas su atsitiktinai pasirinkta Twitter žinute

Žinutė	Pakeitimas	Balai	Nuotaika
High-street #fashion store #Zara accused of using #slave #labour at factories in #Argentina	bazinė žinutė	neigiama: 0.155, neutralu: 0.845, teigiama: 0.0, balas: -0.296	Neigiama 😞
High-street #fashion store #Zara accused of using #slave #labour at factories in #Argentina!!!!!!	su šauktukais	neigiama: 0.219, neutralu: 0.781, teigiama: 0.0, balas: -0.5216	Neigiama 😞
High-street #fashion store #Zara accused of using <u>#SLAVE</u> <u>#LABOUR</u> at factories in #Argentina	panaudojant DIDŽIĄSIAS RAIDES	neigiama: 0.155, neutralu: 0.845, teigiama: 0.0, balas: -0.296	Neigiama 😞

Žinutė	Pakeitimas	Balai	Nuotaika
High-street #fashion store #Zara accused of using #slave #labour at factories in #Argentina but #farewages in Bangladesh	kontrastinis jungtukas – bet	neigiama: 0.091, neutralu: 0.909, teigiama: 0.0, balas: -0.1531	Neigiama 😞
High-street #fashion store #Zara accused of using terrible #slave #labour at factories in #Argentina	laipsnio modifikatoriai (sustiprina nuotaiką)	neigiama: 0.306, neutralu: 0.694, teigiama: 0.0, balas: -0.6486	Neigiama 😞
High-street #fashion store #Zara isn't accused of using #slave #labour at factories in #Argentina	poliškumo keitimas (neigimas)	neigiama: 0.0, neutralu: 0.873, teigiama: 0.127, balas: 0.2235	Teigiama 😊

Iš šios analizės galime daryti išvadą, kad skyrybos ženklų, o ypač šauktukų panaudojimas, laipsnio modifikatorius sustiprina emociją. Pradinė žinutė yra negatyvi taigi šiuo atveju negatyvi nuotaika dar labiau sustiprėja 😞. Poliškumo pakeitimas žinutę verčia į teigiamą 😊, o didžiųjų raidžių panaudojimas įtakos nedaro.

Toliau atlikta analizė su 3 duomenų rinkiniais. Duomenų rinkiniai sudaryti iš *Twitter* žinučių aplinkosaugos, socialinės atsakomybės ir valdymo temų. Kiekviena tema analizuojama atskirai, o žinutės skirstomos 3 klases: teigiamą, neigiamą ir neutralią. Pagrindinės statistikos aplinkosaugos (**E**) tema pateiktos 6 lentelėje, socialinės atsakomybės (**S**) tematika – 7 lentelėje, o valdymo (**G**) tema – 8 lentelėje.

6 lentelė. Žinučių statistiniai duomenys aplinkosaugos (**E**) tema

Žinučių kiekis – 432, unikalių žinučių – 379				
	neigiama	neutralu	teigiama	balas
Mediana	0,0312	0,7904	0,0857	0,1386
Standartinis nuokrypis	0,0778	0,2880	0,1218	0,3549
Min	0	0	0	-0,9092
Max	0,4830	1,0000	0,517	0,9300

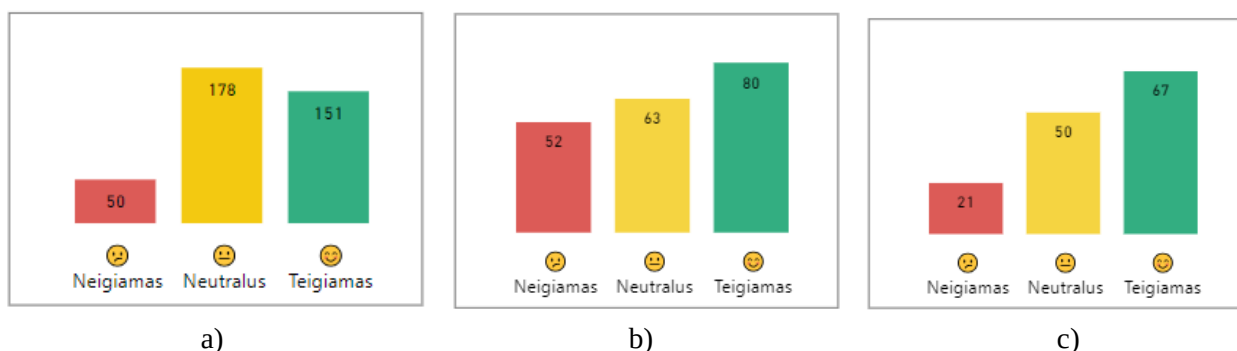
7 lentelė. Žinučių statistiniai duomenys socialinės atsakomybės (**S**) tema

Žinučių kiekis – 207, unikalių žinučių – 195				
	neigiama	neutralu	teigiama	balas
Mediana	0,0881	0,7925	0,1097	0,0410
Standartinis nuokrypis	0,1300	0,1858	0,1321	0,4619
Min	0	0	0	-0,9371
Max	0,6040	1,0000	0,6000	0,9423

8 lentelė. Žinučių statistiniai duomenys valdymo (G) tema

Žinučių kiekis – 139, unikalių žinučių – 138				
	neigiama	neutralu	teigiama	balas
Mediana	0,0438	0,8542	0,1019	0,1664
Standartinis nuokrypis	0,0930	0,1453	0,1118	0,4068
Min	0	0	0	-0,8779
Max	0,4680	1,0000	0,4520	0,9423

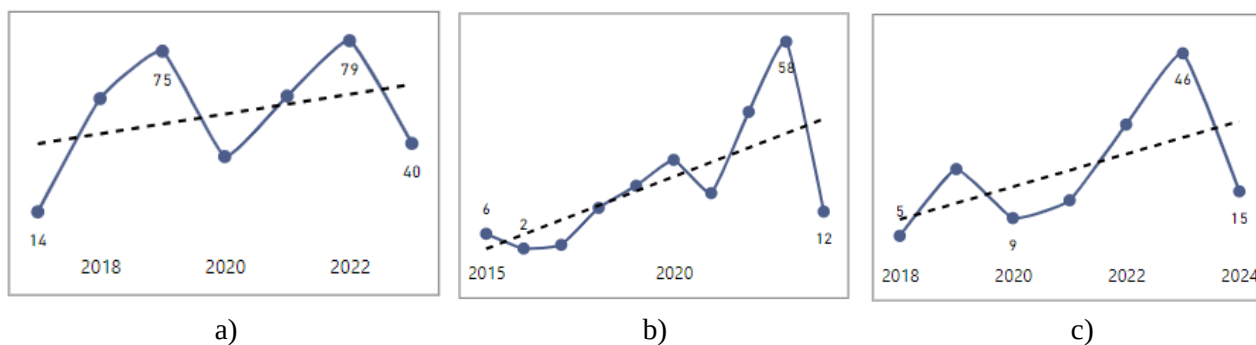
Pagal sudėtinį balą (angl. *Compound*) žinutės paskirstomos į neigiamas, neutralias ir teigiamas, o jų pasiskirstymas pateiktas 30 pav.



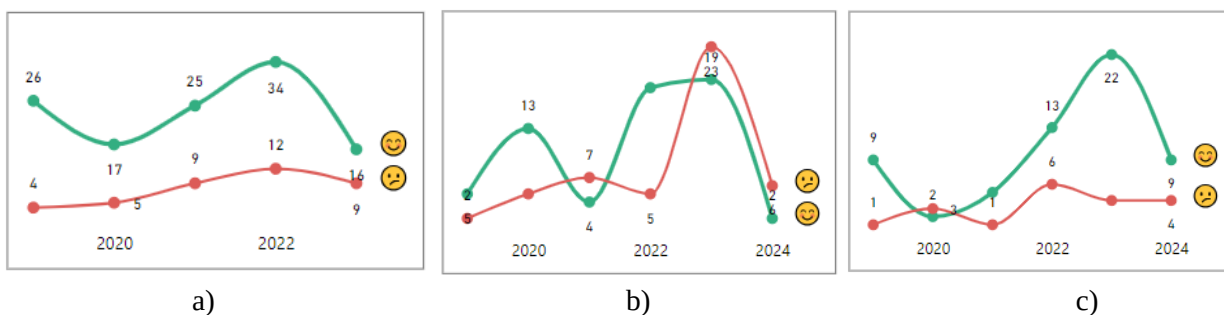
30 pav. Žinučių pasiskirstymas tarp neigiamų, neutralių ir teigiamų skirtingose srityse: a) aplinkosauga, b) socialinė atsakomybė, c) valdymas

Iš 30 pav. vizualiai pateiktų duomenų galime matyti, kad dominuoja neutralios ir teigiamos žinutės. Socialinės atsakomybės srityje žinutės nuotaikos pasiskirstę gana tolygiai, tačiau čia neigiamų nuotaikų pasitaikė daugiausiai, lyginant su kitų temų žinutėmis.

Surinktų žinučių kiekio pasiskirstymą per laikotarpį su trendo linija matome 31 pav., o pastarųjų 5-erių metų žinučių pasiskirstymas tarp teigiamų ir neigiamų atvaizduotas 32 pav.



31 pav. Duomenų pasiskirstymo kiekis su trendo linija: a) aplinkosauga, b) socialinė atsakomybė, c) valdymas



32 pav. 5-erių metų teigiamų ir neigiamų žinučių kiekis: a) aplinkosauga, b) socialinė atsakomybė, c) valdymas

Svarbu atsižvelgti į neigiamas nuotaikas, ar jos yra atsitiktinės, ar pasikartojančios ir stebėti, kaip tai daro įtaką kompanijos reputacijai. Vartotojo sąsajoje pateiktos švieslentės leidžia analizuoti *Twitter* platformoje talpinamų žinučių nuotaikas, bei stebėti pačių žinučių turinį (žr. 2 priedą).

Duomenų tikslumui įvertinti buvo atliktas ekspertinis žinučių vertinimas, priskiriant žinutei vieną iš požymių: teigiamas, neigiamas ar neutralus. Taip įvertintos žinutės palyginamos su VADER metodu įvertintomis žinutėmis ir apskaičiuojamos metrikos kiekvienaia temai atskirai bei pateikiamos 9 lentelėje.

9 lentelė. Žinučių metrikų vertinimas VADER metodu

	Aplinkosauga			Socialinė atsakomybė			Valdymas		
rodiklis / klasė	neig.	neutr.	teig.	neig.	neutr.	teig.	neig.	neutr.	teig.
Pasiskirstymas	21	376	35	91	52	64	32	27	80
Bendrasis tikslumas	56 %			61 %			60 %		
Tikslumas	19 %	94 %	15 %	80 %	61 %	50 %	71 %	37 %	75 %
Jautrumas	48 %	56 %	69 %	52 %	77 %	64 %	47 %	70 %	63 %
F1 balas	27 %	70 %	25 %	63 %	68 %	56 %	57 %	49 %	68 %

Pastebima, kad nepaisant didžiausio žinučių kiekio aplinkosaugos tema, klasės buvo prasčiausiai išskirtos. Taip pat gautas ir prasčiausias bendrojo tikslumo rodiklis, kurie siekia 56 %.

3.3.2. Žinučių analizė atraminių vektorių klasifikatoriumi (SVM)

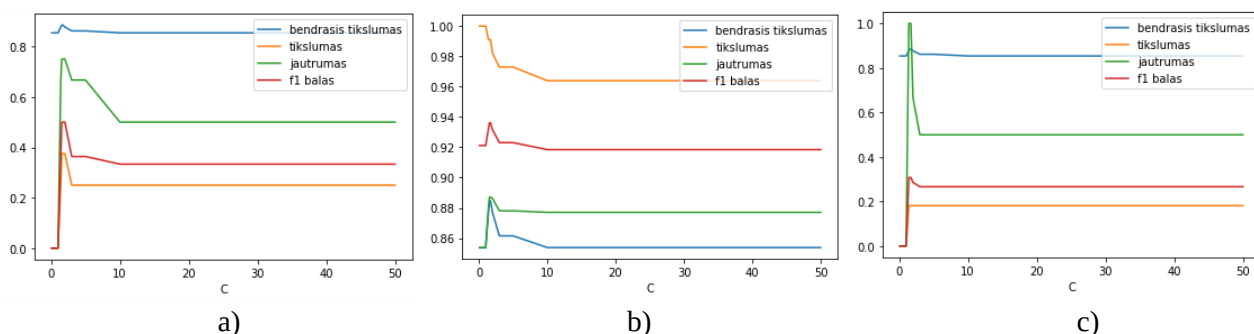
Atraminių vektorių tyrimui naudojama Python *sklearn* biblioteka.

Tyrimas atliktas su trimis duomenų rinkiniais. Kiekvienos temos žinutės (aplinkosauga, socialinė atsakomybė ir valdymas) analizuojamos atskirai. Pradiniai duomenys dalinami į mokymo (70 % pradinių duomenų) ir testavimo (30 % pradinių duomenų) imtis. Turime 3 klases: neigiamą, neutralią, teigiamą.

Siekiant gauti geriausius įverčius, keičiamos C parametro reikšmės. Gauti rezultatai atvaizduojami 10 lentelėje, 11 lentelėje, 12 lentelėje ir 33 pav., 34 pav., 35 pav.

10 lentelė. Žinučių metrikų vertinimas atraminių vektorių metodu aplinkosaugos (*E*) tema

rodiklis / C reikšmės	0,10	1,00	1,55	5,00	10,00	50,00
Pasiskirstymas	[0, 130, 0]	[0, 130, 0]	[4, 124, 2]	[3, 123, 4]	[4, 122, 4]	[4, 122, 4]
Bendrasis tikslumas, %	85	85	88	86	85	85
Tikslumas, %	[0, 100, 0]	[0, 100, 0]	[38, 98, 18]	[25, 97, 18]	[25, 96, 18]	[25, 96, 18]
Jautrumas, %	[0, 85, 0]	[0, 85, 0]	[75, 88, 100]	[66, 88, 50]	[50, 88, 50]	[50, 88, 50]
F1 balas, %	[0, 92, 0]	[0, 92, 0]	[50, 94, 31]	[37, 92, 27]	[33, 92, 27]	[33, 92, 27]



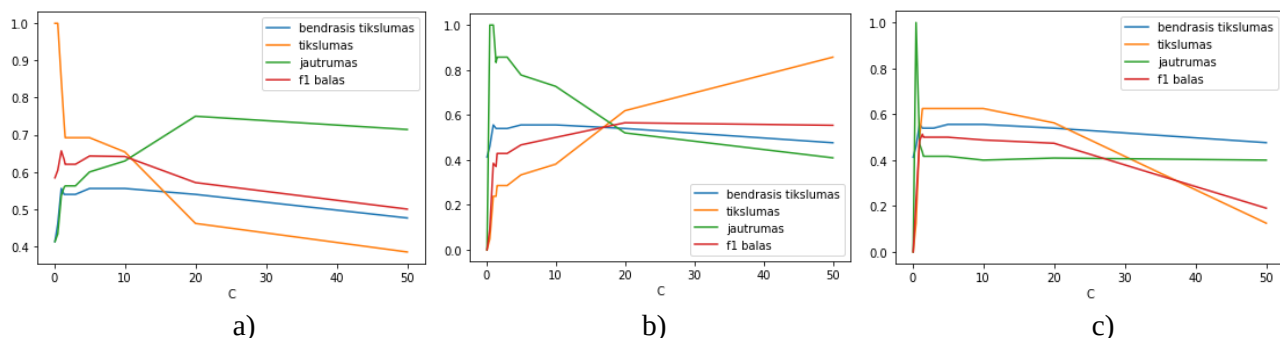
33 pav. Žinučių metrikų vertinimas atraminių vektorių metodu su skirtingomis parametro C reikšmėmis aplinkosaugos (*E*) tema, Oy ašyje – teisingai suklasifikuotų žinučių dalis: a) negatyvios, b) neutralios, c) teigiamos

Geriausias įvertis gautas, kai su mažomis C reikšmėmis klasifikatorius nesugebėjo atskirti klasių, pasiekus reikšmę $C = 1,55$, klasės atsiskyrė ir gautas aukščiausias bendrojo tikslumo vertinimas – 88 %. Toliau didinant C reikšmę, rezultatai įma prastėti.

Analogiškai atlikti bandymai su kitais duomenų rinkiniais. 11 lentelėje pateikti rezultatai su socialinės atsakomybės tema surinktomis žinutėmis.

11 lentelė. Žinučių metrikų vertinimas atraminių vektorių metodu socialinės atsakomybės (*S*) tema

rodiklis / C reikšmės	0,10	1,00	1,55	5,00	10,00	50,00
Pasiskirstymas	[63, 0, 0]	[41, 5, 17]	[32, 7, 24]	[30, 9, 24]	[27, 11, 25]	[14, 44, 5]
Bendrasis tikslumas, %	41	56	54	56	55	48
Tikslumas, %	[100, 0, 0]	[85, 24, 50]	[69, 29, 63]	[69, 33, 63]	[65, 38, 63]	[38, 85, 13]
Jautrumas, %	[41, 0, 0]	[54, 100, 48]	[56, 86, 42]	[60, 77, 42]	[63, 73, 40]	[71, 41, 40]
F1 balas, %	[58, 0, 0]	[66, 38, 49]	[62, 43, 50]	[64, 47, 50]	[64, 50, 49]	[50, 55, 19]



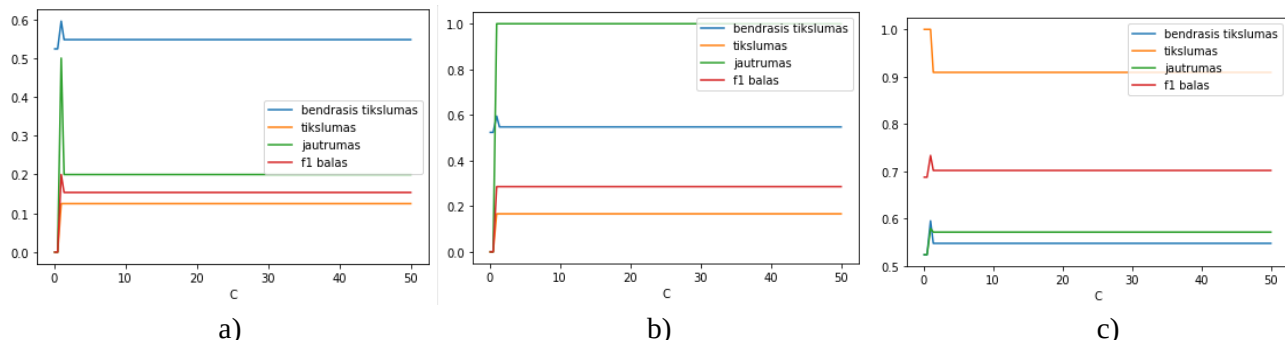
34 pav. Žinučių metrikų vertinimas atraminių vektorių metodu su skirtingomis parametro C reikšmėmis socialinės atsakomybės (S) tema, Oy ašyje – teisingai suklasifikuotų žinučių dalis: a) negatyvios, b) neutralios, c) teigiamos

Šiame duomenų rinkinyje duomenys buvo tolygesni tarp klasių. Geriausias tyrimo rezultatas buvo pasiektas su $C = 5$ reikšme. Tačiau šiuo atveju aukščiausias bendrasis tikslumas siekė tik 56 %.

Trečiojo duomenų rinkinio – žinučių valdymo (G) tema – rezultatai pateikti 12 lentelėje ir 35 pav.

12 lentelė. Žinučių metrikų vertinimas atraminių vektorių metodu valdymo (G) tema

rodiklis / C reikšmės	0,10	1,00	1,55	5,00	10,00	50,00
Pasiskirstymas	[0, 0, 42]	[2, 2, 38]	[5, 2, 35]	[5, 2, 35]	[5, 2, 35]	[5, 2, 35]
Bendrasis tikslumas, %	52	60	55	55	55	48
Tikslumas, %	[0, 0, 100]	[13, 16, 100]	[13, 16, 91]	[13, 16, 91]	[13, 16, 91]	[13, 16, 91]
Jautrumas, %	[0, 0, 52]	[50, 100, 58]	[20, 100, 57]	[20, 100, 57]	[20, 100, 57]	[20, 100, 57]
F1 balas, %	[0, 0, 69]	[20, 29, 73]	[15, 29, 70]	[15, 29, 70]	[15, 29, 70]	[15, 29, 70]



35 pav. Žinučių metrikų vertinimas atraminių vektorių metodu su skirtingomis parametro C reikšmėmis valdymo (G) tema, Oy ašyje – teisingai suklasifikuotų žinučių dalis: a) negatyvios, b) neutralios, c) teigiamos

Šio duomenų rinkinio klasės taip pat buvo nesubalansuotos ir modelis prastai skyrė klases. Su $C = 1$ buvo pasiektas optimaliausias bendrojo tikslumo rezultatas, po kurio rezultatai suprastėjo.

3.3.3. Žinučių analizė Naivaus Bajeso (NB) metodu

Naivaus Bajeso metodas gerai tinkamas kelių klasių klasifikavimo uždaviniams spręsti [80], tačiau šio tyrimo atveju klasifikatorius prastai skyrė duomenis į 3 klases, kai reikėjo klasifikuoti į neigiamą, neutralią ir teigiamą nuotaiką. Žinučių vertinimas Naivaus Bajeso metodu ir metrikos pateiktos 13 lentelėje:

13 lentelė. Žinučių metrikų vertinimas Naivaus Bajeso metodu

	Aplinkosauga			Socialinė atsakomybė			Valdymas		
rodiklis / klasė	neig.	neutr.	teig.	neig.	neutr.	teig.	neig.	neutr.	teig.
Pasiskirstymas	0	130	0	49	2	12	1	0	41
Bendrasis tikslumas	85 %			52 %			55 %		
Tikslumas	0 %	100 %	0 %	92 %	10 %	44 %	12 %	0 %	100 %
Jautrumas	0 %	85 %	0 %	49 %	100 %	58 %	100 %	0 %	54 %
F1 balas	0 %	92 %	0 %	64 %	17 %	50 %	22 %	0 %	70 %

Iš duomenų pasiskirstymo rodiklio matome, kad Naivaus Bajeso metodas žinutes klasifikavo prastai. Tik žinutes socialinės atsakomybės tema atskyrė geriau. Nors bendrasis tikslumas pasirodė gana geras, tačiau beveik visais atvejais žinutes priskyrė vienai klasei.

3.4. Rezultatų vertinimas pritaikius duomenų balansavimo metodus

Analizuojant 3 skirtingus duomenų rinkinius *E*, *S* ir *G* temomis pastebėta, kad duomenų pasiskirstymas į klases yra netolygus, todėl turime nesubalansuotas klases. *S* temai klases buvo geriausiai subalansuotos ir mašininio mokymo metodai neblogai atskyrė duomenis į klases. Tuo tarpu kitiems duomenų rinkiniams pritaikyta duomenų balansavimo technika – imties didinimas (angl. *Random Oversampling*).

3.4.1. VADER sentimentų analizė su subalansuota imtimi

Aplinkosaugos duomenų rinkiniui (*E*) ir valdymo (*G*) duomenų rinkiniui 30 proc. padidinta neigiamų žinučių imtis, kad teigiamų ir neigiamų žinučių imtys būtų maždaug vienodos. Pritaikius imties didinimo techniką ir VADER metodu apskaičiavus tikslumo metrikas, gautos reikšmės pateiktos 14 lentelėje.

14 lentelė. Žinučių metrikų vertinimas VADER metodu pritaikius imties didinimo techniką

	Aplinkosauga			Valdymas		
rodiklis / klasė	neig.	neutr.	teig.	neig.	neutr.	teig.
Pasiskirstymas	125	242	165	72	27	80
Bendrasis tikslumas	72 %			69 %		
Tikslumas	74 %	76 %	65 %	90 %	37 %	75 %
Jautrumas	90 %	70 %	61 %	76 %	70 %	63 %
F1 balas	82 %	73 %	63 %	83 %	49 %	68 %

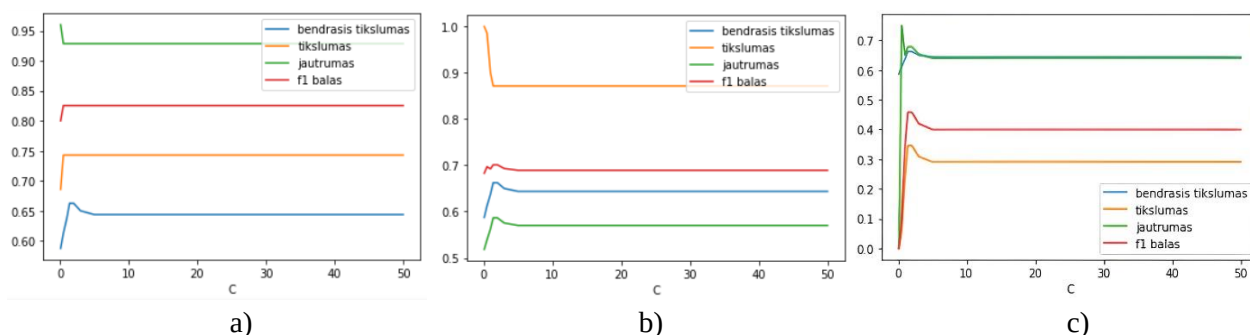
Pasiskirstymas į klases pagerėjo, o tuo pačiu padidėjo ir bendrasis tikslumas abiems duomenų rinkiniams.

3.4.2. Žinučių analizė atraminių vektorių metodu su subalansuota intimi

Analizuojami *E* ir *G* duomenų rinkiniai su aukščiau aptartu imties didinimo metodu. Keičiamos hiperparametro *C* reikšmės siekiant gauti geriausius modelio rezultatus. Gauti rezultatai pateikiami 15 lentelėje ir 36 pav.

15 lentelė. Žinučių metrikų vertinimas atraminių vektorių metodu aplinkosaugos (*E*) tema pritaikius imties didinimo techniką

rodiklis / <i>C</i> reikšmės	0,10	1,00	1,55	5,00	10,00	50,00
Pasiskirstymas	[25, 130, 0]	[28, 112, 4]	[28, 105, 28]	[28, 107, 25]	[28, 107, 25]	[28, 107, 25]
Bendrasis tikslumas, %	59	63	66	64	64	64
Tikslumas, %	[69, 100, 0]	[74, 90, 24]	[74, 87, 35]	[74, 87, 29]	[74, 87, 29]	[74, 87, 29]
Jautrumas, %	[96, 52, 0]	[0, 85, 0]	[93, 58, 67]	[93, 57, 64]	[93, 57, 64]	[93, 57, 64]
F1 balas, %	[80, 68, 0]	[0, 92, 0]	[83, 70, 46]	[83, 69, 40]	[83, 69, 40]	[83, 69, 40]



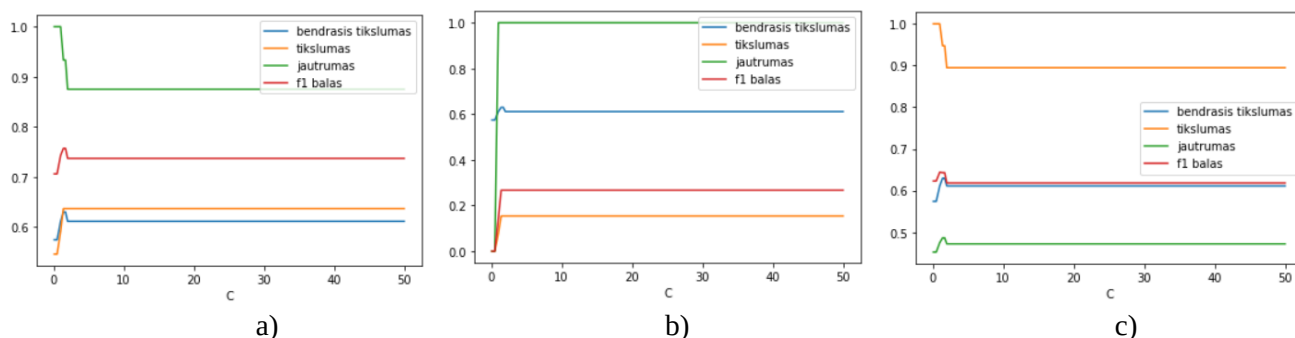
36 pav. Žinučių metrikų vertinimas atraminių vektorių metodu su skirtingomis parametro *C* reikšmėmis aplinkosaugos (*E*) tema, Oy ašyje – teisingai suklasifikuotų žinučių dalis: a) negatyvios, b) neutralios, c) teigiamos

Kaip ir nesubalansuotos imties atveju, geriausia bendrojo tikslumo reikšmė gauta, kai $C = 1,55$. Subalansuotos imties atveju bendrasis tikslumas sumažėjo nuo 88 % iki 66 %, tačiau klasės atsiskyrė daug geriau.

Duomenų rinkiniui socialinės (*S*) atsakomybės tema nebuvo taikoma balansavimo technika, o rezultatai valdymo (*G*) temos žinutėms pateikti 16 lentelėje ir 37 pav.

16 lentelė. Žinučių metrikų vertinimas atraminių vektorių metodu valdymo (*G*) tema pritaikius imties didinimo techniką

rodiklis / <i>C</i> reikšmės	0,10	1,00	1,55	5,00	10,00	50,00
Pasiskirstymas	[12, 0, 42]	[13, 1, 40]	[15, 2, 37]	[16, 2, 36]	[16, 2, 36]	[16, 2, 36]
Bendrasis tikslumas, %	57	61	63	61	61	61
Tikslumas, %	[55, 0, 100]	[59, 8, 100]	[63, 15, 95]	[64, 15, 89]	[64, 15, 89]	[64, 15, 89]
Jautrumas, %	[100, 0, 45]	[100, 100, 48]	[93, 100, 49]	[88, 100, 47]	[88, 100, 47]	[88, 100, 47]
F1 balas, %	[71, 0, 62]	[74, 14, 64]	[76, 27, 64]	[74, 27, 62]	[74, 27, 62]	[74, 27, 62]



37 pav. Žinučių metrikų vertinimas atraminių vektorių metodu su skirtingomis parametro C reikšmėmis valdymo (G) tema, O_y ašyje – teisingai suklasifikuotų žinučių dalis: a) neigatyvios, b) neutralios, c) teigiamos

Atraminių vektorių metodas su G temos žinutėmis geriausius rezultatus išgavo su $C = 1,55$ hiperparametro reikšme. Pritaikius duomenų subalansavimo techniką bendrojo tikslumo rezultatas pakilo iki 63 %.

3.4.3. Žinučių analizė Naivaus Bajeso metodu su subalansuota imtimi

Nesubalansuotoms klasėms Naivaus Bajeso metodas prastai skyrė klases. Subalansavus klases, jos buvo geriau klasifikuotos. Naivaus Bajeso metodu su subalansuota imtimi gauti rezultatai pateikti 17 lentelėje.

17 lentelė. Žinučių metrikų vertinimas Naivaus Bajeso metodu pritaikius imties didinimo techniką

	Aplinkosauga			Valdymas		
rodiklis / klasė	neig.	neutr.	teig.	neig.	neutr.	teig.
Pasiskirstymas	33	118	9	14	0	40
Bendrasis tikslumas	63 %			57 %		
Tikslumas	82 %	93 %	10 %	59 %	0 %	95 %
Jautrumas	88 %	55 %	67 %	93 %	0 %	45 %
F1 balas	85 %	69 %	19 %	72 %	0 %	61 %

Atlikus analizę su nesubalansuota imtimi, Naivaus Bajeso metodu visos žinutės buvo priskirtos vienai klasei. Kai žinutės yra subalansuotos, šis metodas geriau atskiria klases. Ir nors bendrasis tikslumas sumažėjo (aplinkosaugos žinučių atveju) arba nežymiai pakito (valdymo žinučių atveju), visi kiti rodikliai pagerėjo.

3.5. Tyrimo rezultatų apžvalga

Klasifikavimas paremtas VADER metodu yra paprasčiausias iš visų nagrinėtų algoritmų ir jam atlikti nereikalinga atskirti testavimo imties. Tačiau šiame metode nėra galimybių, leidžiančių pagerinti algoritmo rezultatus. Tekstui priskiriamas poliariškumo vertinimas, kuris vėliau paverčiamas į sudėtinį balą (angl. *Compound*). Pagal šio balo reikšmes, tekstą priskiriame vienai ar kitai klasei. VADER metodo rezultatai pagerėjo subalansavus imtį. Šiuo metodu atliktas vertinimas pateiktas 18 lentelėje.

18 lentelė. Apibendrintas žinučių metrikų vertinimas VADER metodu

	E	E – subalansuota imtis	S	G	G – subalansuota imtis
Pasiskirstymas	[21, 376, 35]	[125, 242, 165]	[91, 52, 64]	[32, 27, 80]	[72, 27, 80]
Bendrasis tikslumas, %	56	72	61	60	69
Tikslumas, %	[19, 94, 15]	[74, 76, 65]	[80, 61, 50]	[71, 37, 75]	[90, 37, 75]
Jautrumas, %	[48, 56, 69]	[90, 70, 61]	[52, 77, 64]	[47, 70, 63]	[76, 70, 63]
F1 balas, %	[27, 70, 25]	[82, 73, 63]	[63, 68, 56]	[57, 49, 68]	[83, 49, 68]

Iš 18 lentelės duomenų matyti, kad tiek E, tiek G temos žinutėms duomenų balansavimas davė teigiamus rezultatus ir pagerino klasių atskyrimą bei bendrąjį tikslumą.

Atraminių vektorių metodas (SVM) reikalavo skirstyti duomenis į mokymo ir testavimo imtis. Šis metodas pasirodė universalesnis, nes keičiant C parametro reikšmes buvo gauti optimalūs rezultatai. Atraminių vektorių metodo rezultatai taip pat pagerėjo subalansavus duomenų imtį. Geriausi rezultatai pateikti 19 lentelėje.

19 lentelė. Geriausių rezultatų palyginimas gautas atraminių vektorių metodu

	E	E – subalansuota imtis	S	G	G – subalansuota imtis
C reikšmė	1,55	1,55	5,00	1,00	1,55
Pasiskirstymas	[4, 124, 2]	[28, 105, 28]	[30, 9, 24]	[2, 2, 38]	[15, 2, 37]
Bendrasis tikslumas, %	88	66	56	60	63
Tikslumas, %	[38, 98, 18]	[74, 87, 35]	[69, 33, 63]	[13, 16, 100]	[63, 15, 95]
Jautrumas, %	[75, 88, 100]	[93, 58, 67]	[60, 77, 42]	[50, 100, 58]	[93, 100, 49]
F1 balas, %	[50, 94, 31]	[83, 70, 46]	[64, 47, 50]	[20, 29, 73]	[76, 27, 64]

Atraminių vektorių metodu gauti rezultatai svyruoja nuo 56 % iki 66 % (subalansuotos imties atveju) ir yra prastesni nei klasifikuojant VADER metodu, bet geresni nei Naivaus Bajeso atveju.

Naivaus Bajeso (NB) metodas silpnai skyrė duomenų imtį į klases. Šis metodas gana paprastas naudotis, tačiau nebuvo galimybės keisti parametrų norint optimizuoti rezultatus. Apibendrinti rezultatai pateikiami 20 lentelėje.

20 lentelė. Apibendrintas žinučių metrikų vertinimas Naivaus Bajeso metodu

	E	E – subalansuota imtis	S	G	G – subalansuota imtis
Pasiskirstymas	[0,130, 0]	[33, 118, 9]	[49, 2, 12]	[1, 0, 41]	[14, 0, 40]
Bendrasis tikslumas, %	85	63	52	55	57
Tikslumas, %	[0, 100, 0]	[82, 93, 10]	[92, 10, 44]	[12, 0, 100]	[59, 0, 95]
Jautrumas, %	[0, 85, 0]	[88, 55, 67]	[49, 100, 58]	[100, 0, 54]	[93, 0, 45]
F1 balas, %	[0, 92, 0]	[85, 69, 19]	[64, 17, 50]	[22, 0, 70]	[72, 0, 61]

Subalansavus imtį Naivaus Bajeso metodo bendrasis tikslumas pagerėjo analizuojant valdymo temos (**G**) žinutes ir pakilo iki 57 %, o aplinkosaugos (**E**) duomenų rinkiniui pagerėjo atskyrimas į klases.

Anksčiau aptartame P. Sudhir’as [48] tyrime taip pat buvo panaudotas atraminių vektorių metodas ir Naivaus Bajeso metodas sentimentų analizei. Naivaus Bajeso metodas buvo gautas 54,77 % bendrasis tikslumas, o atraminių vektorių metodu – 88,30 %. Aukštesni rezultatai buvo gauti taikant gilaus mokymo (LSTM, GRU) ir transformerių modelių metodus (BERT, BERT-UDA).

Tiriamajame projekte panaudoti įprastiniai mašinimo mokymo algoritmai – VADER, atraminių vektorių ir Naivaus Bajeso. Geriausi rezultatai pasiekti VADER metodu, toliau – atraminių vektorių klasifikatoriumi. Prasčiausi rezultatai gauti Naivaus Bajeso metodu.

Išanalizavus surinktus duomenis pastebėta, kad *Twitter* daugiau pasitaikė teigiamo pobūdžio žinučių, negu neigiamų. Gana didelis kiekis žinučių buvo informacinio pobūdžio, neturinčios intensyvaus sentimentų (teigiamo arba neigiamo) ir dėl to pateko į neutralią klasę bei nedavė didelės naudos tyrimui. Pastebėta, kad visuomenė gana palankiai vertina kompanijų pastangas skirtas aplinkosaugai ir valdymo skaidrumo temoms. Negatyvių žinučių pasitaikė daugiau kalbant apie socialinę atsakomybę.

Teigiamai vertintos žinutės skelbė apie inovacijas, apdovanojimus, taip pat ten, kur paminėti pozityvūs žodžiai: *great, possitive, smart, most valuable, friendly* ir pan. Žinutės su žodžiais *waste, recycling, labour, wages, workers right, finance, corruption* dažnu atveju buvo vertintos kaip neigiamos.

Kompanijos, norinčios didinti savo reputaciją ir būti patrauklios darbuotojams bei investuotojams turėtų atsakingai skelbti ESG informaciją, o taip pat atsižvelgti į visuomenės nuomonę. Sentimentų analizė – tinkamas būdas stebėti visuomenės nuomonę ir greitai į ją reaguoti.

3.6. Rekomendacijos

Tyrimas buvo atliktas su specialiai šiam tyrimui surinktais duomenimis – *Twitter* žinutėmis. Tyrimo eigai pagerinti siūlomos rekomendacijos, kurios pateikiamos 21 lentelėje.

21 lentelė. Iššūkiai ir pasiūlymai įvairiems sentimentų analizės etapams

Etapas	Iššūkis	Pasiūlymas
Duomenų surinkimas	Dėl <i>Twitter</i> duomenų politikos apribojimų tapo sudėtingas žinučių surinkimas.	Rasti automatizuotų būdų išgauti <i>Twitter</i> žinutes arba naudoti kitą duomenų šaltinį viešajai nuomonei tirti.
Duomenų apdorojimas	Duomenys nestructūruoti ir dėl to tekstas gerai neatsiskyrė pagal aprašytą logiką	Pritaikius automatizuotą duomenų surinkimo metodą, būtų surinkti structūruoti duomenys
Požymių išgavimas	Parinkti tinkamus metodus požymiams išgauti	Nurodyti konkrečius stop žodžius, neišmesti neiginių, nes jie gali pakeisti sakinio poliškumą
Žinučių vertinimas	Norint įvertinti žinutes, reikalingos ekspertinės žinios	Žinutes paskirti vertinti keliems ekspertams
Modelio parinkimas	Išsirinkti tinkamiausią algoritmą iš galimų variantų	Išbandyti papildomus mašinio mokymo metodus, kuris geriausiai tiktų turimiems duomenims
Rezultatų vertinimas	Parinkti tinkamus hiperparametrus, įvertinti klasių disbalansą	Panaudoti klasių balansavimo technikas, keisti hiperparametrus ir vertinti modelio patikimumą keletu rodiklių
Vartotojo sąsaja	Parengti standartinę, lengvai adaptuojamą verslui švieslentę	Pasigilinti į konkretaus vartotojo poreikius ir adaptuoti švieslentę

21 lentelėje pateikiami visi tyrimo eigos etapai, iššūkiai, su kuriais buvo susidurta vykdant tyrimą, bei rekomendacijos, kaip patobulinti gautus rezultatus ir pritaikyti tyrimą platesniu mastu. Atliktas tyrimas skirtas *Twitter* žinučių analizei, tačiau jį adaptavus gali būti pritaikytas ir kitam nestructūruotui tekstui analizuoti.

Kompanijos, paskelbę aplinkosaugos, socialinės ir valdymo informaciją įsipareigoja visuomenei, darbuotojams bei investuotojams, todėl negali atsiriboti nuo visuomenės ir turi aktyviai sekti socialinėse platformose vyraujančias nuomones. Itin svarbu tinkamai reaguoti į neigiamas žinutes, stebėti, kas išprovokuoja neigiamas reakcijas, kokie teiginiai dominuoja. Pagal šią informaciją atitinkamai turi būti parengtos ir įmonių komunikacinės kampanijos.

Kompanijoms sukurtos švieslentės leidžia lengvai sekti aktualią informaciją, greitai reaguoti į vyraujančias nuotaikas ir priimti tinkamus sprendimus. Vartotojo sąsaja leidžia bet kokiam vartotojui, neturinčiam programavimo patirties matyti rezultatus, o atlikus papildomų konfigūravimų, galima gauti žinučių sentimentų analizę realiu laiku.

Išvados

1. Aplinkosaugos, socialinės ir valdymo (ESG) tema pasaulyje tampa vis aktualesnė. Visuomenė nori atvirumo ir skaidrumo, investuotojai siekia sužinoti daugiau nei tik finansinius rezultatus, o esamiems ar potencialiems darbuotojams svarbios įmonių vertybės bei darbo sąlygos. Kompanijos, norinčios didinti savo reputaciją ir būti patrauklios darbuotojams bei investuotojams turėtų atsakingai skelbti ESG informaciją, o taip pat atsižvelgti į visuomenės nuomonę socialinėje žiniasklaidoje. ESG informacijos atskleidimas svarbus investicijų pritraukimui bei rizikos valdymui, didina įmonės reputaciją ir skaidrumą, užtikrina reguliavimo atitiktį laikymąsi, pritraukia jaunus darbuotojus, kuriems svarbi socialinė ir aplinkosauginė atsakomybė. Nuo 2024 m. įsigaliojusi įmonių tvarumo ataskaitų teikimo direktyva CSRD dar labiau įpareigos kompanijas atsižvelgti į poveikį aplinkai, socialinę atsakomybę bei skaidrų įmonių valdymą.
2. Informacinės technologijos ir interneto plėtra suteikia galimybę greitai ir plačiai skleisti informaciją, o socialinė žiniasklaida tampa pagrindiniu kanalu, per kurį visuomenė gali išreikšti savo nuomonę. Tuo tarpu mados pramonė yra labai jautri visuomenės nuomonei, stipriai priklauso nuo vartotojų įpročių bei elgsenos, todėl kompanijoms labai svarbu suprasti vartotojų lūkesčius, nuotaikas ir turėti prieinamą būdą stebėti šiuos pokyčius. *Twitter* – vienas populiariausių sklaidos kanalų, kur per dieną sugeneruojama apie 500 mln. žinučių, todėl yra puikus duomenų šaltinis sentimentų analizei. Atsižvelgiant į didėjančią susirūpinimą aplinkos tarša, darbo sąlygomis mados pramonėje, įmonės vis dažniau skelbia apie savo ESG iniciatyvas, siekdamos ne tik patenkinti reguliavimo reikalavimus, bet ir gerinti savo įvaizdį vartotojų akyse.
3. Sentimentų analizei atlikti tinkami mašininio mokymo metodai, tokie kaip Naivus Bajesas (NB), logistinė regresija (LogR), atraminių vektorių metodas (SVM), atsitiktiniai miškai (RF), sprendimų medžiai (DT), o taip pat giliojo mokymo algoritmai: sąsūkų neuroniniai tinklai (CNN), rekurentiniai neuroniniai tinklai (RNN), LSTM. Giliojo mokymosi metodams mokyti reikalingos didelės apimties duomenų imtys, todėl tyrimui atlikti pasirinkti mašininio mokymo metodai dėl turimos santykinai mažos duomenų apimties.
Sentimentų analizės tyrimui panaudotas VADER, Naivaus Bajeso ir atraminių vektorių metodai. Paprasčiausias iš šių metodų – VADER, nes jis nereikalauja atskirti mokymo imties duomenų ir yra grįstas leksikonu ir taisyklėmis. Šio metodo bendrasis tikslumas priklauso nuo duomenų rinkinio skyrėsi, bet siekė 72 %. Naivaus Bajeso metodo tikslumas siekė iki 83 % bendrojo tikslumo, tačiau tik vienam duomenų rinkiniui, kai klases atskyrė prastai. Kitais atvejais Naivaus Bajeso metodo bendrasis tikslumas svyravo 55-63 % ribose, bet žemesnio rezultato buvo galima tikėtis išanalizavus mokslinėje literatūroje skelbiamus tyrimus. Atraminių vektorių metodu buvo pasiektas 88 % bendrasis tikslumas. Dauguma atvejų duomenų tikslumas buvo geresnis analizuojant subalansuotas imties duomenis.
4. Išanalizavus surinktas *Twitter* žinutes, susijusias su mados pramonės kompanijomis pastebėta, kad daugiau pasitaikė teigiamo pobūdžio žinučių, negu neigiamų. Taip pat didelis kiekis žinučių pateko į neutralią klasę, kas tyrimui didelės naudos nedavė. Neutralaus pobūdžio žinutės atsirado dėl žinutės prasto informatyvumo, informacinio pobūdžio teksto arba netinkamo teksto atskyrimo. Galima daryti išvadą, kad visuomenė gana palankiai vertina kompanijų pastangas skirtas aplinkosaugai ir valdymo skaidrumo temoms. Negatyvių žinučių daugiau pasitaikė susijusių su socialine atsakomybe ir darbuotojų išnaudojimu. Į šią informaciją įmonėms reikėtų atkreipti didesnę dėmesį, kuriant įmonių strategijas ir išklaidyti susidarytą įvaizdį.
5. Vartotojo sąsajai sukurti panaudotas Microsoft Power BI analitikos įrankis. Sukurta švieslėntė leidžia lengvai analizuoti *Twitter* žinutes – pastebėti kokiomis temomis daugiausiai kalbama,

kokie raktiniai žodžiai pavartoti, kokios nuotaikos dominuoja kalbant konkrečia (**E**, **S** ar **G**) tema. Taip pat galime matyti žinučių turinį, išskirti blogiausiai įvertintas žinutes, o esat poreikiui – modifikuoti švieslentę pridedant papildomos informacijos. Švieslentė yra interaktyvi ir ją galima pasiekti pasirinkus šią nuorodą: [ESG sentimentų analizės švieslentė](#)

Literatūros sąrašas

1. ESG – aplinkos apsauga, socialinė atsakomybė ir valdysena [interaktyvus]. [žiūrėta 2024-02-08]. Prieiga per: <https://tvarumoklausimai.lt/esg>.
2. LOUREIRO, M. L., ALLÓ, M. and COELLO, P. Hot in Twitter: Assessing the emotional impacts of wildfires with sentiment analysis. *Ecological Economics* [interaktyvus]. 2022, 200, 107502. ISSN 0921-8009. Prieiga per: <https://doi.org/10.1016/j.ecolecon.2022.107502>.
3. NI, Z. Corporate ESG Investment: Factors Affecting the Economic Sustainability. *Advances in Economics, Management and Political Sciences* [interaktyvus]. 2023, 39(1), 210–214 [žiūrėta 2024-02-03]. ISSN 2754-1177. Prieiga per: doi:10.54254/2754-1169/39/20231971.
4. PARGENDLER, M. The Rise of International Corporate Law [interaktyvus]. *European Corporate Governance Institute – Law Working Paper*, 2020 [žiūrėta 2024-02-10]. Prieiga per: <http://dx.doi.org/10.2139/ssrn.3728650>.
5. RUGGIE J. G., The United Nations and Globalization: Patterns and Limits of Institutional Adaptation [interaktyvus]. *Global Governance* 9, p. 301-321, 2003 [žiūrėta 2024-02-10]. Prieiga per: [https://scholar.harvard.edu/files/john-ruggie/files/the_united_nations_and_globalization_patterns_and_limits_of_institution_0_0.pdf](https://scholar.harvard.edu/files/john-ruggie/files/the_united_nations_and_globalization_patterns_and_limits_of_institutional_adaptation_0_0.pdf).
6. THÉRIEN, J. P. and POULIOT V. *The Global Compact: Shifting the Politics of International Development?* *Global Governance: A Review of Multilateralism and International Organizations* [interaktyvus]. 2006, 12(1), 55–76 [žiūrėta 2024-02-10]. ISSN 1942-6720. Prieiga per: doi:10.1163/19426720-01201006.
7. KELL, G. Relationships of International Organizations with Other Actors, Ch.34 Relations with the Private Sector. Iš: *The Oxford Handbook of International Organizations* [interaktyvus]. Oxford University Press, 2016 [žiūrėta 2024-02-17]. ISBN 9780199672202. Prieiga per: doi:10.1093/law/9780199672202.003.0034.
8. United Nations. Secretary-General Proposes Global Compact on Human Rights, Labour, Environment, in Address to the World Economic Forum in Davos, 1999 [žiūrėta 2024-02-17]. Prieiga per: <https://press.un.org/en/1999/19990201.sgsm6881.html>.
9. World bank. Who Cares Wins, 2004-08. IFC Advisory Services in Environmental and Social Sustainability [interaktyvus]. [žiūrėta 2024-02-17]. Prieiga per: <https://documents1.worldbank.org/curated/en/444801491483640669/pdf/113850-BRI-IFC-Breif-whocares-PUBLIC.pdf>.
10. CARLSON D. Mentions of ESG and Sustainability are Being Made on Thousands of Corporate Earnings Calls [Interaktyvus]. Iš: MarketWatch, 2021-07-19 [žiūrėta 2024-02-03]. Prieiga per: <https://www.marketwatch.com/story/mentions-of-esg-and-sustainability-are-being-made-on-thousands-of-corporate-earnings-calls-11626712848>.
11. STEDMAN C. ESG strategy and management: Complete guide for business [Interaktyvus]. 2023 [žiūrėta 2024-02-03]. Prieiga per: <https://www.techtarget.com/sustainability/feature/ESG-strategy-and-management-Complete-guide-for-businesses>.
12. ROUEN E., SACHDEVA K. and YOON A. The Evolution of ESG Reports and the Role of Voluntary Standards. *SSRN Electronic Journal* [Interaktyvus]. 2022 [žiūrėta 2024-03-10]. ISSN 1556-5068. Prieiga per: doi:10.2139/ssrn.4227934.
13. Google trends [žiūrėta 2024-02-03]. Prieiga per: <https://trends.google.com>.

14. MANDAS M., et al. ESG in the financial industry: What matters for rating analysts? *Research in International Business and Finance* [interaktyvus]. 2023, 66, 102045 [žiūrėta 2024-03-17]. ISSN 0275-5319. Prieiga per: doi:10.1016/j.ribaf.2023.102045.
15. POLLMAN E. The Making and Meaning of ESG. 2022 [interaktyvus]. U of Penn, Inst for Law & Econ Research Paper No. 22-23, European Corporate Governance Institute – Law Working Paper No. 659/2022, Harvard Business Law Review, Forthcoming, Prieiga per: SSRN <https://ssrn.com/abstract=4219857>.
16. LUTKEVICH B. What does the “G” in ESG mean? Publikuota 2023-12-04 [žiūrėta 2024-03-01]. Prieiga per: <https://www.techtarget.com/whatis/feature/What-does-the-G-in-ESG-mean>.
17. WOOD D. et al. What Do We Mean by the S in ESG? Society as a Stakeholder in Responsible Investment 553. Iš: *The Routledge Handbook of Responsible Investment*. 2015.
18. CURTIS Q. et al. Do ESG Mutual Funds Deliver on Their Promises? 2021 [interaktyvus]. Michigan Law Review, Forthcoming, U of Penn, Inst for Law & Econ Research Paper No. 21-17, European Corporate Governance Institute – Law Working Paper No. 586/2021, Prieiga per: SSRN: <https://ssrn.com/abstract=3839785>.
19. <https://scrapfly.io/blog/how-to-scrape-twitter/> [žiūrėta 2024-03-02].
20. Twitter Statistics For 2024: Users, Demographics, Trends. 2023 [žiūrėta 2024-03-02]. Prieiga per: <https://whatsthebigdata.com/twitter-statistics/>.
21. YUE L. et al. A survey of sentiment analysis in social media. *Knowledge and Information Systems* [interaktyvus]. 2018, 60(2), 617–663 [žiūrėta 2024-03-24]. ISSN 0219-3116. Prieiga per: doi:10.1007/s10115-018-1236-4.
22. ANGER I. and KITTL C. Measuring influence on Twitter. Iš: *the 11th International Conference* [interaktyvus]. New York, New York, USA: ACM Press, 2011 [žiūrėta 2024-03-11]. ISBN 9781450307321. Prieiga per: doi:10.1145/2024288.2024326.
23. LTP Annual Sustainability Performance Report 2021. [žiūrėta 2024-03-11]. Prieiga per: <https://publications.ltpgroup.com/view/1010610611/i/>.
24. REMY N., SPEELMAN E. and SWARTZ S. Style that’s sustainable: A new fast-fashion formula. McKinsey. 2016 [žiūrėta 2024-03-11]. Prieiga per: https://www.mckinsey.com/capabilities/sustainability/our-insights/style-thats-sustainable-a-new-fast-fashion-formula#.
25. The True Cost. A documentary film exploring the impact of fashion on people and the planet that was released in 2015. [žiūrėta 2024-03-11]. Prieiga per: <https://truecostmovie.com/>.
26. Facts & Key figures 2022 of the European Textile and Clothing Industry; Euratex, 2022. [žiūrėta 2024-03-11]. Prieiga per: https://euratex.eu/wp-content/uploads/EURATEX_FactsKey_Figures_2022rev-1.pdf.
27. Fashion E-commerce Global Market Report 2024, Research and Markets, 2024 [žiūrėta 2024-03-11]. Prieiga per: <https://www.researchandmarkets.com/reports/5939666/fashion-e-commerce-global-market-report>.
28. Global Clothing B2C E-Commerce Market 2021, Research and Markets, 2021 [žiūrėta 2024-03-11]. Prieiga per: <https://www.researchandmarkets.com/reports/5509676/global-clothing-b2c-e-commerce-market-2021>.

29. Patagonia. Environmental responsibility. Prieiga per:
<https://www.patagonia.com/environmental-responsibility-materials/>.
30. HENISZ W., KOLLER T. and NUTTALL R. Five ways that ESG creates value. McKinsey Quarterly 4 (2019): 1-12 [žiūrėta 2024-03-22]. Prieiga per:
<https://www.mckinsey.com/~media/McKinsey/Business%20Functions/Strategy%20and%20Corporate%20Finance/Our%20Insights/Five%20ways%20that%20ESG%20creates%20value/Five-ways-that-ESG-creates-value.ashx>.
31. DAGILIENĖ L., BRUNECKIENĖ J., VARANIŪTĖ V., BANIONIENĖ J. Circular Business Models in the Manufacturing Industry. Springer International Publishing, 2023 [žiūrėta 2024-04-04]. ISBN 9783031288081. Prieiga per: doi:10.1007/978-3-031-28809-8.
32. HUNTER B. Dokumentinis filmas. 2022 [žiūrėta 2024-03-22]. Prieiga per:
<https://www.fashionreimaginedfilm.com/>.
33. YU H., AHN M., and HAN E. Key driver of textile and apparel industry management: fashion brand ESG and brand reputation. *Frontiers in Environmental Science* [interaktyvus]. 2023 [žiūrėta 2024-04-04]. ISSN 2296-665X. Prieiga per: doi:10.3389/fenvs.2023.1140004.
34. JIA F., YIN S., CHEN L. and CHEN X. The circular economy in the textile and apparel industry: A systematic literature review. *Journal of Cleaner Production* [interaktyvus]. 2020, 259, 120728 [žiūrėta 2024-04-05]. ISSN 0959-6526. Prieiga per: doi:10.1016/j.jclepro.2020.120728.
35. European Parliament. The impact of textile production and waste on the environment. 2024 [žiūrėta 2024-03-23]. Prieiga per:
<https://www.europarl.europa.eu/topics/en/article/20201208STO93327/the-impact-of-textile-production-and-waste-on-the-environment-infographics>.
36. International Labour Organization. ILO adopts code of practice on safety and health in textiles, clothing, leather and footwear industries. 2021 [žiūrėta 2024-03-23]. Prieiga per:
https://www.ilo.org/global/about-the-ilo/newsroom/news/WCMS_822368/lang--en/index.htm.
37. Ellen Macarthur Foundation. Fashion and the circular economy – deep dive. [žiūrėta 2024-03-23]. Prieiga per: <https://www.ellenmacarthurfoundation.org/fashion-and-the-circular-economy-deep-dive>.
38. ADEGEEST D. A. Garment workers among the lowest-paid industrial workers globally. 2024 [žiūrėta 2024-04-05]. Prieiga per: <https://fashionunited.com/news/fashion/garment-workers-among-the-lowest-paid-industrial-workers-globally/2024010557687>.
39. BANERJEE A. 20 Certifications and Standards for Textile Industry Businesses. Online Clothing Study. 2024 [žiūrėta 2024-04-10]. Prieiga per:
<https://www.onlineclothingstudy.com/2022/03/20-popular-certifications-and-standards.html>.
40. Textile Certifications. Your guide in choosing certifications. [žiūrėta 2024-04-10]. Prieiga per:
https://www.vestilanatura.it/en/textile-certifications/?gad_source=1&gclid=Cj0KCQjw-mvBhDwARIsAA-Q0Q6ZeRxX8lhaMQ--u6b8hzaRAUWAEzWfAKgX52TUg0tH_hLWOnV6WxgaAqghEALw_wcB.
41. ENGEL J. F, BLACKWELL R. D. & MINIARD P. W. Consumer behavior. Hinsdale, IL, USA: The Dryden Press, 1990. ISBN: 0030229790.
42. MOSLER H. J. et al. Deriving Interventions on the Basis of Factors Influencing Behavioral Intentions for Waste Recycling, Composting, and Reuse in Cuba. *Environment and*

- Behavior* [interaktyvus]. 2008, 40(4), 522–544 [žiūrėta 2024-04-10]. ISSN 1552-390X. Prieiga per: doi:10.1177/0013916507300114.
43. KWON S. H. A study on circular textiles in luxury fashion. *J. Commun. Des.* 2022, 78, 84–101. I410-ECN-0102-2023-600-000554290.
 44. DRUS Z. and KHALID H. Sentiment Analysis in Social Media and Its Application: Systematic Literature Review. *Procedia Computer Science* [interaktyvus]. 2019, 161, 707–714 [žiūrėta 2024-04-10]. ISSN 1877-0509. Prieiga per: doi:10.1016/j.procs.2019.11.174.
 45. DEVIKA, M. D., SUNITHA C., and GANESH A. Sentiment Analysis: A Comparative Study on Different Approaches. *Procedia Computer Science* [interaktyvus]. 2016, 87, 44–49 [žiūrėta 2024-04-12]. ISSN 1877-0509. Prieiga per: doi:10.1016/j.procs.2016.05.124.
 46. JAGDALE R. S., SHIRSAT V. S. and DESHMUKH S. N. Sentiment Analysis on Product Reviews Using Machine Learning Techniques. Iš: *Cognitive Informatics and Soft Computing* [interaktyvus]. Singapore: Springer Singapore, 2018, pp. 639–647 [žiūrėta 2024-04-13]. ISBN 9789811306167. Prieiga per: doi:10.1007/978-981-13-0617-4_61.
 47. What is natural language processing (NLP)? [žiūrėta 2024-04-13]. Prieiga per: <https://www.ibm.com/topics/natural-language-processing>.
 48. SUDHIR P. and SURESH V. D. Comparative Study of Various Approaches, Applications and Classifiers for Sentiment Analysis. *Global Transitions Proceedings* [interaktyvus]. 2021 [žiūrėta 2024-04-13]. ISSN 2666-285X. Prieiga per: doi:10.1016/j.gltp.2021.08.004.
 49. WANKHADE M., RAO A. C. S. and KULKARNI C. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review* [interaktyvus]. 2022 [žiūrėta 2024-04-13]. ISSN 1573-7462. Prieiga per: doi:10.1007/s10462-022-10144-1.
 50. SÁNCHEZ-RADA, J. F. and IGLESIAS C. A. Social context in sentiment analysis: Formal definition, overview of current trends and framework for comparison. *Information Fusion* [interaktyvus]. 2019, 52, 344–356 [žiūrėta 2024-04-14]. ISSN 1566-2535. Prieiga per: doi:10.1016/j.inffus.2019.05.003.
 51. SINGH N. and JAISWAL U. C. A Detailed Sentiment Analysis Survey Based on Machine Learning Techniques. *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal* [interaktyvus]. 2023, 12(1), e29105 [žiūrėta 2024-04-13]. ISSN 2255-2863. Prieiga per: doi:10.14201/adcaij.29105.
 52. Sentiment Analysis: A Definitive Guide. Monkey Learn. [žiūrėta 2024-03-24]. Prieiga per: <https://monkeylearn.com/sentiment-analysis/>.
 53. FERRARI A. and ESULI A. An NLP approach for cross-domain ambiguity detection in requirements engineering. *Automated Software Engineering* [interaktyvus]. 2019, 26(3), 559–598 [žiūrėta 2024-04-13]. ISSN 1573-7535. Prieiga per: doi:10.1007/s10515-019-00261-7.
 54. <https://brand24.com/> .
 55. <https://sentic.net/> .
 56. CAMBRIA E., ZHANG X., MAO R., CHEN M., KWOK K. SenticNet 8: Fusing emotion AI and commonsense AI for interpretable, trustworthy, and explainable affective computing. 2024 [žiūrėta 2024-04-13]. Iš: HCII. <https://sentic.net/senticnet-8.pdf>.
 57. MEDHAT W., HASSAN A., and KORASHY H. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal* [interaktyvus]. 2014, 5(4), 1093–1113 [žiūrėta 2024-04-12]]. ISSN 2090-4479. Prieiga per: doi:10.1016/j.asej.2014.04.011.
 58. ARORA S. Sentiment Analysis Using Python. *Analytics Vidhya*. 2024 [žiūrėta 2024-04-01].

- Prieiga per: <https://www.analyticsvidhya.com/blog/2022/07/sentiment-analysis-using-python/>.
59. AHUJA R. et al. The Impact of Features Extraction on the Sentiment Analysis. *Procedia Computer Science* [interaktyvus]. 2019, 152, 341–348 [žiūrėta 2024-03-30]. ISSN 1877-0509. Prieiga per: doi:10.1016/j.procs.2019.05.008.
 60. KIRITCHENKO, S., X. ZHU, and S. M. MOHAMMAD. Sentiment Analysis of Short Informal Texts. *Journal of Artificial Intelligence Research* [interaktyvus]. 2014, 50, 723–762 [žiūrėta 2024-03-30]. ISSN 1076-9757. Prieiga per: doi:10.1613/jair.4272.
 61. ZHAO Y. Y., QIN B. and LIU T. Integrating Intra- and Inter-document Evidences for Improving Sentence Sentiment Classification. *Acta Automatica Sinica* [interaktyvus]. 2010, 36(10), 1417–1425 [žiūrėta 2024-03-30]. ISSN 1874-1029. Prieiga per: doi:10.1016/s1874-1029(09)600574.
 62. MOREO A. et al. Lexicon-based Comments-oriented News Sentiment Analyzer system. *Expert Systems with Applications* [interaktyvus]. 2012, 39(10), 9166–9180 [žiūrėta 2024-03-30]. ISSN 0957-4174. Prieiga per: doi:10.1016/j.eswa.2012.02.057.
 63. GEETHA L. Vader: A Comprehensive Guide to Sentiment Analysis in Python. Medium. 2023. [žiūrėta 2024-04-01]. Prieiga per: <https://medium.com/@rslavanyageetha/vader-a-comprehensive-guide-to-sentiment-analysis-in-python-c4f1868b0d2e>.
 64. ZHAO J., LIU K. and XU L. Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. *Computational Linguistics* [interaktyvus]. 2016, 42(3), 595–598 [žiūrėta 2024-04-02]. ISSN 1530-9312. Prieiga per: doi:10.1162/coli_r_00259.
 65. LIU B. Sentiment Analysis and Subjectivity. *Handbook of Natural Language Processing*. 2010 [žiūrėta 2024-04-12]. <https://www.researchgate.net/publication/228667268>.
 66. SHAROU K. A., LI Z., SPECIA L. Towards a Better Understanding of Noise in Natural Language Processing. 2021. Language and Multimodal AI (LAMA) Lab, Imperial College London, UK. Prieiga per: <https://aclanthology.org/2021.ranlp-1.7.pdf>.
 67. SUBRAMANIAM L. V. et al. A survey of types of text noise and techniques to handle noisy text. Iš: *The Third Workshop* [interaktyvus]. New York, New York, USA: ACM Press, 2009 [žiūrėta 2024-03-30]. ISBN 9781605584966. Prieiga per: doi:10.1145/1568296.1568315.
 68. MICHEL P. and NEUBIG G. 2018. MTNT: Atestbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 543–553, Brussels, Belgium. Association for Computational Linguistics. Prieiga per: <https://aclanthology.org/D18-1050.pdf>.
 69. IVANOVA I. Twitter is now X. Here's what that means. MoneyWatch. 2023 [žiūrėta 2024-04-07]. Prieiga per: <https://www.cbsnews.com/news/twitter-rebrand-x-name-change-elon-musk-what-it-means/>.
 70. ANTONAKAKI D., FRAGOPOULOU P. and IOANNIDIS S. A survey of Twitter research: Data model, graph structure, sentiment analysis and attacks. *Expert Systems with Applications* [interaktyvus]. 2021, 164, 114006 [žiūrėta 2024-04-10]. ISSN 0957-4174. Prieiga per: doi:10.1016/j.eswa.2020.114006.
 71. Top 100 companies. [žiūrėta 2024-03-11]. Prieiga per: <https://fashionunited.com/i/top100>.
 72. SWARNKAR N. VADER Sentiment Analysis: A Complete Guide, Algo Trading and More. 2020. [žiūrėta 2024-04-13]. Prieiga per internetą: <https://blog.quantinsti.com/vader-sentiment/>.
 73. NAYAK A., NATARAJAN S. Comparative study of Naïve Bayes, Support Vector Machine and Random Forest Classifiers in Sentiment Analysis of Twitter feeds. *International Journal of Advanced Studies in Computer Science and Engineering IJASCSE Volume 5, Issue 1*, 2016.

- Prieiga per: [https://www.ijascse.org/volume-5-theme-based-issue-1/Support vector machine.pdf](https://www.ijascse.org/volume-5-theme-based-issue-1/Support%20vector%20machine.pdf).
74. What is Parquet? [žiūrėta 2024-04-20]. Prieiga per: <https://www.databricks.com/glossary/what-is-parquet>.
 75. BORG A. and BOLDT M. Using VADER sentiment and SVM for predicting customer response sentiment. *Expert Systems with Applications* [interaktyvus]. 2020, 162, 113746 [žiūrėta 2024-03-30]. ISSN 0957-4174. Prieiga per: doi:10.1016/j.eswa.2020.113746.
 76. GUDINAVIČIENĖ A. CSRD (Corporate Sustainability Reporting Directive) – įmonių tvarumo ataskaitų teikimo direktyva. Valdymo koordinavimo centras. [žiūrėta 2024-04-24]. Prieiga per: https://governance.lt/wp-content/uploads/2023/02/CSRD_02-09_VKC_pranesimas.pdf.
 77. MANGURI H., RAMADHAN N. and MOHAMMED P. R. A. Twitter Sentiment Analysis on Worldwide COVID-19 Outbreaks. *Kurdistan Journal of Applied Research* [interaktyvus]. 2020, 54–65 [žiūrėta 2024-04-13]. ISSN 2411-7706. Prieiga per: doi:10.24017/covid.8.
 78. LYU J. C., HAN E. and LULI G. K. COVID-19 Vaccine–Related Discussion on Twitter: Topic Modeling and Sentiment Analysis. *Journal of Medical Internet Research* [interaktyvus]. 2021, 23(6), e24435 [žiūrėta 2024-04-13]. ISSN 1438-8871. Prieiga per: doi:10.2196/24435.
 79. Accuracy, precision, and recall in multi-class classification. Classification Metrics Guide. Prieiga per: <https://www.evidentlyai.com/classification-metrics/multi-class-metrics>.
 80. Naive Bayes Algorithm. 2021 [žiūrėta 2024-05-01]. Prieiga per: <https://kdagiit.medium.com/naive-bayes-algorithm-4b8b990c7319>.
 81. WANG B. et al. An Empirical Study on Data Balancing in Machine Learning Based Software Traceability Methods. Iš: *2023 International Joint Conference on Neural Networks (IJCNN)* [interaktyvus]. IEEE, 2023 [žiūrėta 2024-04-27]. Prieiga per: doi:10.1109/ijcnn54540.2023.10191386.
 82. MAZUMDER S. 5 Techniques to Handle Imbalanced Data For a Classification Problem. 2023 [žiūrėta 2024-05-05]. Prieiga per: <https://www.analyticsvidhya.com/blog/2021/06/5-techniques-to-handle-imbalanced-data-for-a-classification-problem/>.
 83. PECORELLI F. et al. A large empirical assessment of the role of data balancing in machine-learning-based code smell detection. *Journal of Systems and Software* [interaktyvus]. 2020, 169, 110693 [žiūrėta 2024-05-05]. ISSN 0164-1212. Prieiga per: doi:10.1016/j.jss.2020.110693.

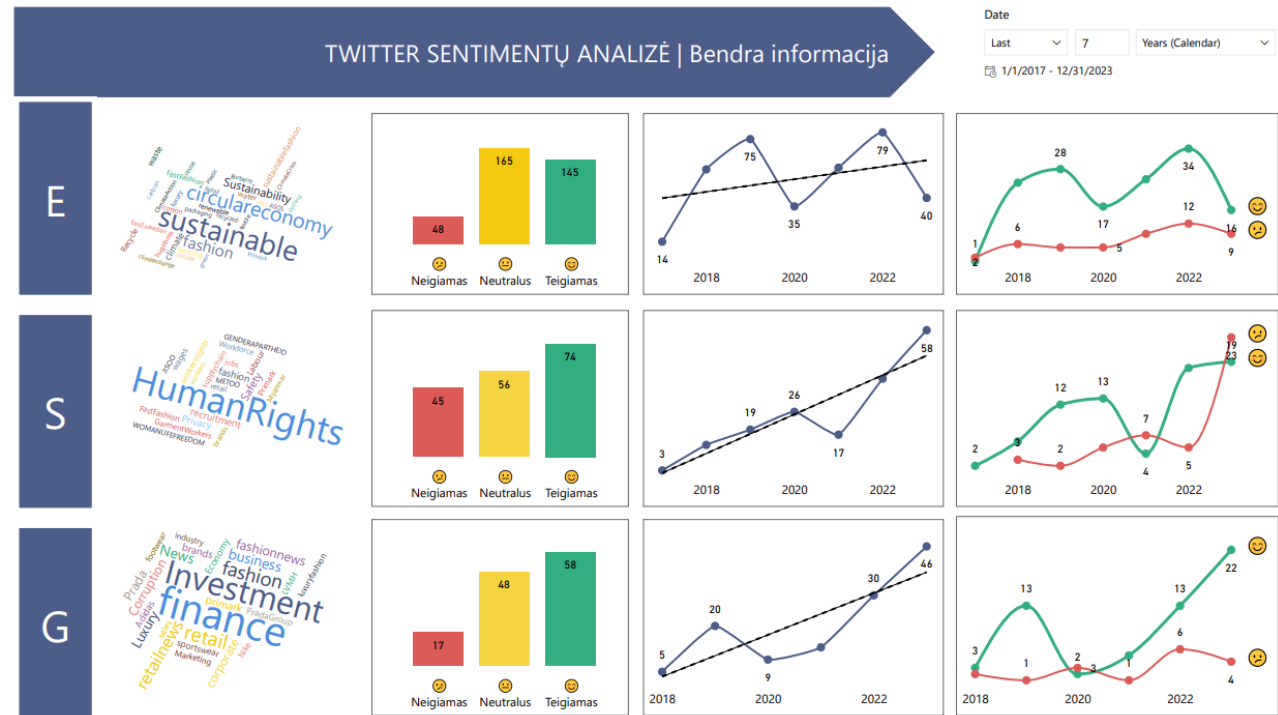
Priedai

1 priedas. Dažniausiai paminėti ESG terminai [14]

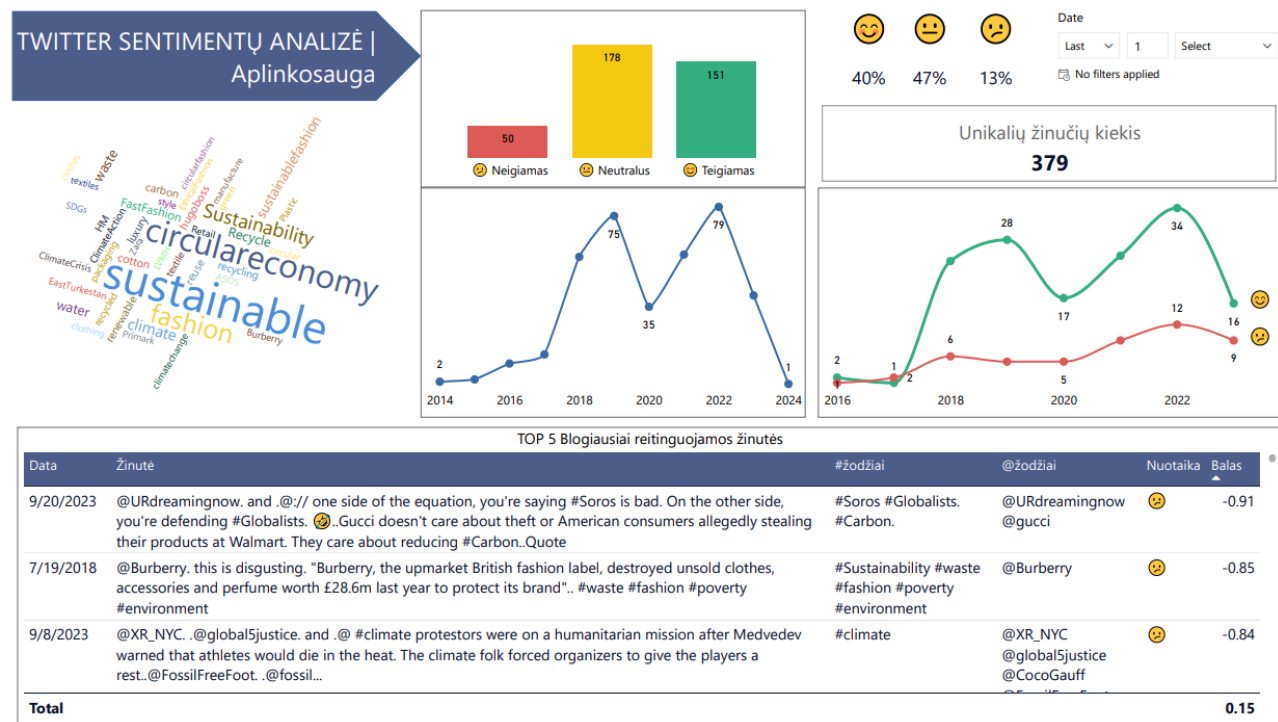
#	ESG	Label and most frequent words
1.	E	Sustainable energy <i>Carbon, Energy, Emission, GHG, Reduce, Regulation, Renewable, Fuel, Power, Gas</i>
2.	E	Environmental and climate change risk <i>Emission, Waste, Community, Effluent, Environmental, Air, Project, Relation, Mine, Spill</i>
3.	E	Sustainable supply chain <i>Water, Supply, Resource, Right, Chain, Supplier, Material, Human, Production, Manufacture</i>
4.	S	Human capital management <i>Human Capital, Diversity, Manage, Initiative, Support, Workforce, Disclose, Growth, Recruitment, Skill</i>
5.	S	Human capital empowerment <i>Human Capital, Labour, Development, Initiative, Turnover, Train, Talent, Diversity, Rate, Develop, Experience</i>
6.	S	Data privacy and security <i>Data, Security, Privacy, Information, Breach, Process, Cybersecurity, Sensitive, Financial, Cyberattacks</i>
7.	S	Consumer protection <i>Quality, Governance, Safety, Standard, Recall, Test, Ensure, Certification, Food, ISO</i>
8.	S	Health and safety <i>Safety, Health, Occupational, Fatality, Incident, Contractor, Train, Accident, Performance, Injury</i>
9.	G	Business ethics and code of conduct <i>Ethic, Corruption, Bribery, Ethical, Whistleblower, Anticompetitive, Bank, Fine, Tax, Launder</i>
10.	G	Product governance transparency <i>Disclose, Conduct, Major, Commitment, Access, Potential, Train, Lack, Initiative, Strategy</i>
11.	G	ESG integration in investment and financing decision <i>Integration, Investment, Financials, Bank, Responsible, Opportunity, Asset, Project, Diversify, Finance</i>
12.	G	Real estate green finance <i>Build, Property, Real Estate, Revenue, Tenant, Green, Safety, Portfolio, Climate, Building</i>
13.	G	Ethics conduct and misselling <i>Governance, Quality, Financial, Client, Bank, Offer, Lend, Control, Reputational, Poor</i>

2 priedas. Vartotojo sąsaja parengta pagal VADER sentimentų analizės metodą Power BI analitikos įrankiu

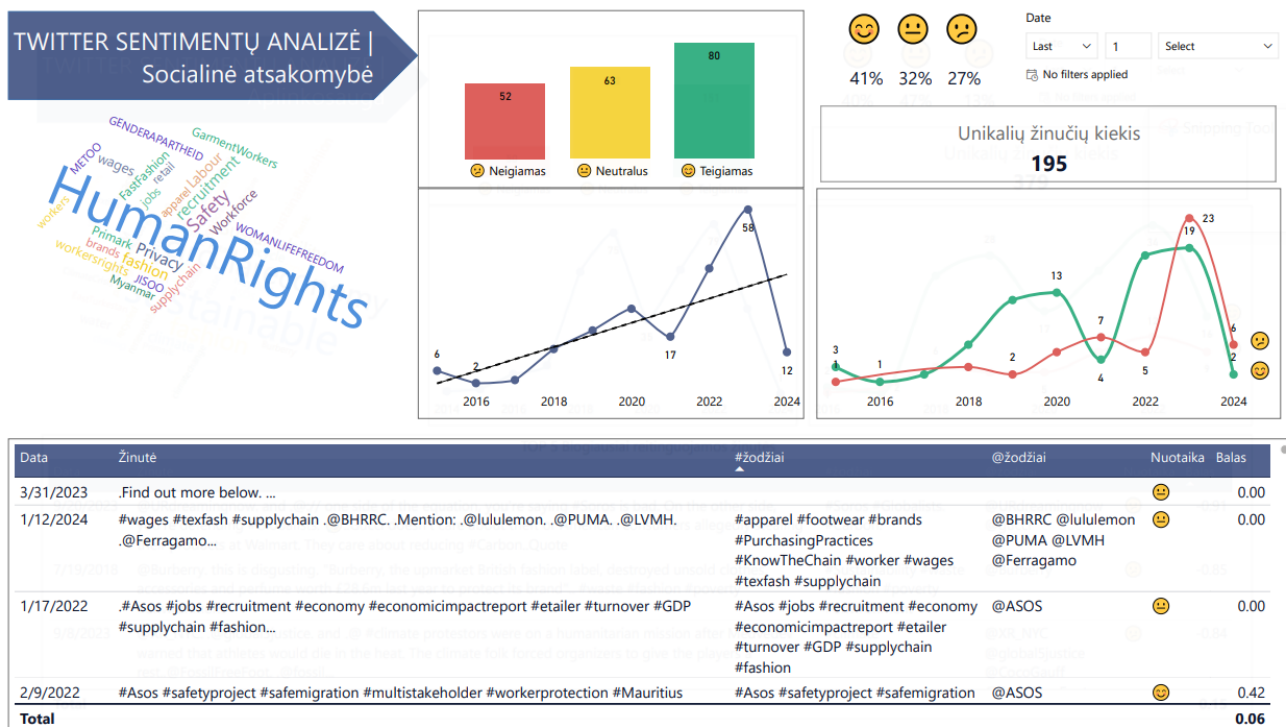
Nuoroda į interaktyvią vartotojo sąsają:
ESG sentimentų analizės švieslentė



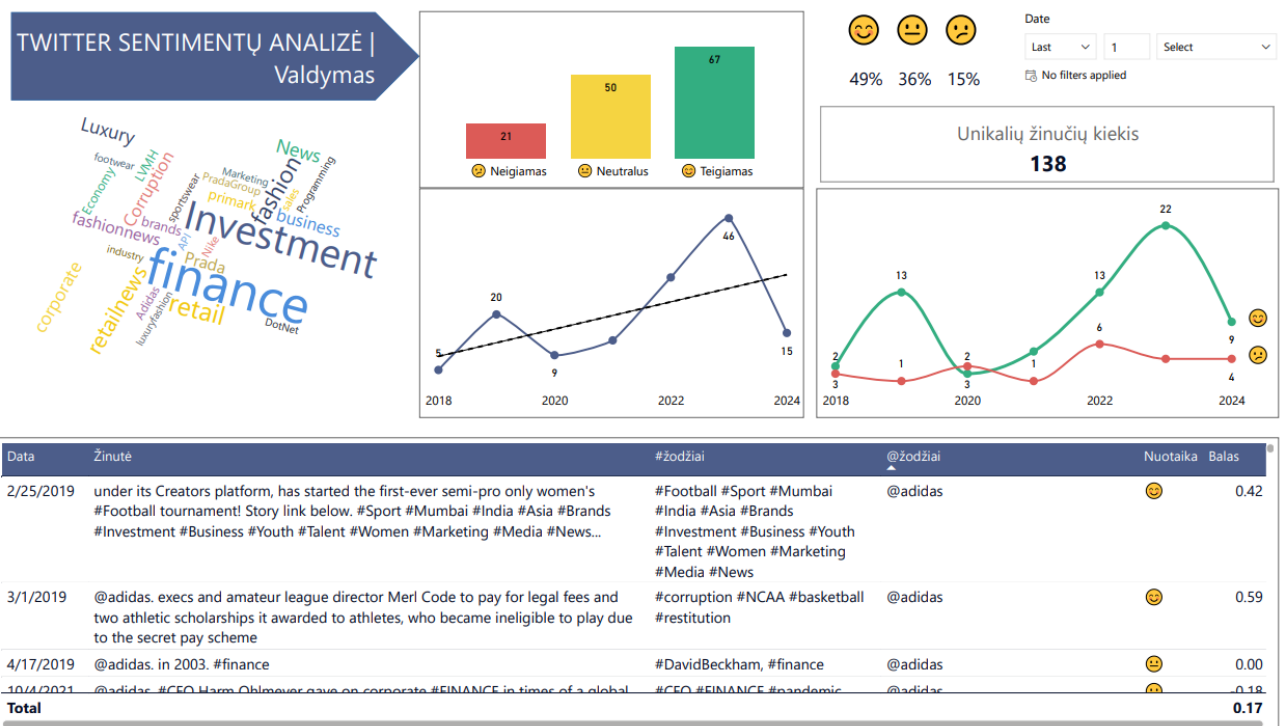
1 pav. Švieslentės pavyzdys. Twitter sentimentų analizė: Bendra informacija ESG tema



2 pav. Švieslentės pavyzdys. Twitter sentimentų analizė: sentimentai anlinkosaugos (E) tema



3 pav. Švieslentės pavyzdys. Twitter sentimentų analizė: sentimentai socialinės atsakomybės (S) tema



4 pav. Švieslentės pavyzdys. Twitter sentimentų analizė: sentimentai valdymo (G) tema