VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
INFORMATIKOS INSTITUTAS

Baigiamasis magistro darbas

Socialiniuose tinkluose naudojamų emotikonų ir sentimentų analizės vertinimo tyrimas

Atliko:
Justas Tamašauskas parašas

Vadovas:
dr. Linas Bukauskas

Vilnius
2018

Turinys

Anotacija

Šiomis dienomis beveik kiekvienas žmogus naudoja bent vieną socialinę platformą ar dalyvauja kokiame nors internetiniame forume. Visi šie portalai yra milžiniški tekstu išreikštų emocijų, nuomonių bei argumentų, duomenų rinkiniai – įvairių sričių analitikų svajonė atliekant sentimentinę analizę. Šiame darbe buvo siekiama sukurti ir įgyvendinti sentimentiniu žodynu pagrįstą semantinės orientacijos (SO) algoritmą *SentiII* ir išbandyti jo veikimą analizuojant „Facebook“ socialiniuose tinkluose publikuojamus tekstus. Šis algoritmas įvertina sakinių SO įverčio susietumo priklausomybę bei tekste esančius kontekstinius keitiklius (pavyzdžiui, inversijas, skyrybos ženklus, stiprinančius / silpninančius žodelius ir t.t.). Vėliau *SentiII* buvo papildytas sukurtais 10 emotikonų modeliais, kurie rėmėsi „Facebook“ reakcijų atlikta analize. Algoritmo testavimas buvo atliktas pateikus 150 tekstus iš populiarių lietuviškų „Facebook“ grupių lyginant jų SO įverčius su *SentiII* suteiktomis vertėmis.

Summary

Research of Evaluation Analysis When Compering Social Network Emojis and Sentiment Analysis

Now since almost every person has access to social network or hub, expression of opinion, statement or emotion became more visible. Such data set is a dream for the businesses and various analysts. It enables to measure impact of ad campaigns, success of new product or simply understand reaction to news or selected communication strategy. However, people tend to express opinions in complex ways. Rhetorical devices like sarcasm, irony, and implied meaning can mislead, thus complicating sentiment analysis task.

In this work lexicon based semantic orientation (SO) algorithm *SentiII* was created which takes into account SO of past sentences and various contextual converters. By analyzing texts from popular English and Lithuanian „Facebook“ pages, multiple strategies were tested how should previous text's sentences a ect ongoing SO score. Also good practices were listed how to handle various contextual elements (e.g. inversions, intensifiers or syntax). Moreover 10 semantic models were introduced (based on „Facebook“ reactions) which later extended SO algorithm. *SentiII* was tested using 150 texts from popular Lithuanian „Facebook“ pages comparing it with expert evaluations.

Įvadas

Aplinkinių žmonių požiūris, nuomonė apie naujienas, gyvenimo aspektus ar kasdienius dalykus visais laikais darė įtaką mūsų pasirinkimams ar net bendram požiūriui į įvarius aspektus. Prieš atsirandant interneto kultūrai, žmonių požiūrį formavo šeimos nariai, draugai ar koks nors žinomas visuomenės veikėjas. Atsiradus internetui ir jo architektūriniam atnaujinimui, WEB 2.0, žmonės pradėjo dalintis savo pažiūromis ir nuomonėmis be jokių ribų: tinklaraščiuose, forumų diskusijose bei komentarų skiltyse, elektroniniuose naujienų portaluose, socialiniuose tinkluose. Nors šiais laikais internetiniuose portaluose savo nuomonę dažniausiai reiškia mums nepažįstami žmonės, kurie nėra profesionalūs kritikai ar žurnalistai, jų požiūris neretai tampa pavyzdžiu, kuriuo remdamiesi darome sprendimus.
Anot 2016 metų atliktos 1062 JAV gyventojų apklausos [8]:

- 84% žmonių pasitiki internetiniais atsiliepimais lygiai taip pat, kaip ir asmeninėmis rekomendacijomis;
- 7 iš 10 žmonių paliktų atsiliepimą ar komentarą apie naują produktą/paslaugą, jei jie būtų paprašyti;
- 90% asmenų perskaito mažiausiai 10 atsiliepimų prieš formuodami savo nuomonę.

Gebėjimas apdoroti tokią įvairovę nuomonių, atsiliepimų yra vienas iš esminių verslo įrankių suprasti subjektyvias priežastis, kaip ir kodėl klientas reaguoja į produktus ar paslaugas [39]. Pavyzdžiui, ar naujovė yra sėkminga? Ką klientai mano apie siūlomas paslaugas? Ar klientų aptarnavimo skyrius patenkina lūkesčius? Kokius šalutinius efektus sukelia siūlomo produkto naudojimas? [20, 9] Tokie duomenys taip pat yra plačiai naudojami politikos mokslams [36], sociologijos raidos tyrimams [17, 14] arba bendrų tendencijų įvertinimui [15, 35]. Šis nuomonių tyrimo uždavinys dar vadinamas sentimentine teksto analize, kurios tikslas yra nustatyti teksto kontekstinį poliariškumą ar emocinį įvertį. Kai per 1 sekundę sukuriama 10 naujų svetainių, parašomas 51 naujas tinklaraštis ir 54000 naujų įrašų socialiniame tinkle „Facebook“ [7], galima suprasti, kad vienkartinės ar ekspromtinės tokio pobūdžio analizės greitai taps nebeaktualios ir neatspindinčios besikeičiančios realybės [22]. Tampa akivaizdu, kad reikalingas autonominis sprendimas, kuris ne tik apdorotų didelius tekstų kiekius, tačiau gebėtų nustatyti atsiliepimų autorių tekstuose perteiktas emocijas, jas apibendrintų ir periodiškai atnaujintų.

Šio uždavinio įgyvendinimas tapo daug lengvesnis, kai 2016 metų pradžioje didžiausias socialinis tinklas „Facebook“ pristatė 5 naujus reakcijos emotikonus šalia įprastos „patinka“ (angl. *Like*) opcijos. Dabar vartotojas gali pasirinkti: „nustebęs“ (angl. *WOW*), „piktas“ (angl. *Angry*), „besijuokiantis“ (angl. *HaHa*), „mylintis“ (angl. *Love*), „liūdnas“ (angl. *Sad*) ir tą patį „patinka“ emotikonus, taip galėdamas konkrečiau išreikšti savo emociją.

Darbo tikslas yra sukurti emocinio lygio įvertinimo modelį šiuolaikiniams socialiniuose tinkluose publikuojamiems lietuvių kalbos tekstams ir jį patobulinti emotikonų pridedama informacija. Siekiant įgyvendinti įvardytą tikslą buvo išsikelti šie **uždaviniai**:

1. Sukurti ir įgyvendinti sentimentų vertinimo modelį;
2. Įvertinti modelio veikimo tikslumą, bei optimalią nustatymų konfigūraciją.
3. Ištirti kaip „Facebook“ portaluose naudojami emotikonai reprezentuoja teksto sentimentinę informaciją.

4. Palyginti sukurto sentimentų vertinimo modelio gaunamus įverčius su emotikonų perteikiama informacija.

5. Patobulinti sukurtą modelį pagal 3 užduotyje gautus rezultatus.

Išsikeltam tikslui ir uždaviniams pasiekti buvo apžvelgta jau atliktų tyrimų metodika bei rezultatai. Remiantis esamomis metodikomis ir kitų darbų išvadomis buvo pilnai suprogramuotas sentimentinės analizės algoritmas (jis įvardintas *SentiII*) Python kalba, nenaudojant jokių sentimentų analizės bibliotekų. Sukurto algoritmo veikimas paremtas lingvistinių taisyklių ir emocijų žodyno modelio taikymu (kuris taip pat buvo sugeneruotas pagal emocijų vertinimo teorijas), bei „Facebook“ emotikonų teikiama sentimentine informacija. Kelios pagrindinės įgyvendintos lingvistinės taisyklės bei ypatybės algoritme yra:

- sakinio poliariškumo nustatymas (įvertinant inversiją, stiprinimo / silpninimo žodelius bei sakinio struktūros faktorius);
- atsižvelgimas į tekstą sudarančių sakinių emocinį susietumą;
- teksto ir sakinių emocinio lygio kitimo vertinimas.

Atlikto tyrimo metu *SentiII* algoritmas buvo testuojamas pateikiant socialiniame tinkle „Facebook“ publikuojamus tekstus iš kelių populiariausių lietuviškų puslapių: „Laikykitės Ten“, „Delfi.lt“, „Geriausios Demotyvacijos“, „Remigijus Šimašius“ bei „Andrius Kubilius“. Kadangi šių puslapių moderatoriai yra grasūs visuomenės veikėjai ar viešos / privačios institucijos, jų pranešimai yra pasiekiami plačiai visuomenės daliai bei prižiūrimi profesionalų, išlaikant gramatiškai taisyklingą ir stilistiškai tvarkingą tekstą. Tokiu būdu tai yra puikus šaltinis atliekant semantinės orientacijos įverčių tyrimą. Siekiant įvertinti algoritmo veikimo kokybę bei identifikuoti spragas, nustatytas tekstų emocinis lygis buvo lyginamas pirmą kartą su tiesioginiu žmogaus teksto įverčiu, o antrą – su emotikonų reprezentuojama reakcija.

Šiuo tiriamuoju darbu yra sprendžiamos kelios semantinės orientacijos įverčio uždavinio problemos. Visų pirma, naudojama lietuvių kalba pasižymi sudėtinga morfologija bei sintakse, o tai pat jos vartojimas internete dažnai nėra taisyklingas struktūriškai. Įvairios meninės raiškos priemonės naudojamos tekste taip pat komplikuoja šį uždavinį. Į šių dienų aktualijas ir požiūrius į tam tikrus aspektus taip pat turi būti atsižvelgta. Dėl šių priežasčių algoritmas turi turėti tam tikras neuniversalias išimtis, o taip pat ir pagalbinius šaltinius. Žvelgiant bendrai į sentimentų analizės metodologijas, pasigendama aiškios sakinių emocinio susietumo strategijos, o egzistuojantys būdai siūlo tiesiogiai sumuoti sakinių įverčius, taip ignoruojant praeitų sakinių emocinius svorius.

Šis darbas remiasi ankstesniu atliktu mokslo tiriamuoju darbu [31]. 1 skyriuje aptariamas emocinio įverčio uždavinys ir galimi sprendimo būdai, įvardijami sentimentų žodyno pagrįsto sprendimo pranašumai bei „Facebook“ emotikonų panaudojimas. 2 skyriuje aprašomas *SentiII* algoritmo veikimo modelis, o 3 skyriuje aptariami pasirinkti algoritmo parametrai. 4 skyriuje aprašoma emotikonų kuriama sentimentinė informacija, kuria pasinaudojus 5 skyriuje buvo apibendrinti emotikonų modeliai bei pateikti *SentiII* algoritmo rezultatai su ir be emotikonų modelio papildymo. 6 ir 7 skyriuose pateikiamos išvados ir tolimesnių tyrimų gairės.

1. Sentimentų analizės uždavinio apžvalga

Pagrindinis emocinio įverčio uždavinys yra nustatyti, kaip autoriaus emocija yra perteikta tekste ir ar ji yra neigiamo, ar teigiamo poliariškumo. Taigi, emocinis įvertis apima:

- sentimentų analizę – tai yra bendrinis metodas gauti teksto poliariškumą;
- semantinį orientacijos (angl. *semantic orientation, SO*) nustatymą – žodžių, sakinių ar teksto poliariškumo stiprumo įvertinimas.

Egzistuoja du pagrindiniai būdai, kaip yra automatizuojami sentimentų analizės uždaviniai. Pirmas, tai žodynais pagrįstas metodas, kai semantinė teksto orientacija vertinama pagal sakinio žodžius ir jų nešamą emocinį įvertį [37, 18]. Antras, tai teksto klasifikavimo metodas, kartais įvardijamas kaip sistemos mokymosi (angl. *machine-learning*) metodas, pagrįstas sukurta klasifikavimo sistema naudojant jau pažymėtus tektus ar sakinius [25]. Šiame tyrime buvo dirbama naudojant sentimentinį žodyną, kuriame kiekvienam žodžiui buvo priskirtas poliariškumas (teigiamas, neigiamas) bei semantinė orientacija (emocinis įvertis nuo -3 iki 3).

1.1. Sentimentų žodynu pagrįstas metodas

Semantiniai žodynai gali būti generuojami dvejais metodais, renkant ir dedant žodžius rankiniu būdu [34] arba identifikuojant daiginius žodžius (angl. *seed words*) ir taip adaptyviai plečiant žodynus [38, 27]. Dauguma žodyno pagrindu veikiančių semantinės analizės algoritmų fokusuojasi į būdvardžius kaip į raktinį indikatorių nustatant sakinių semantinę orientaciją [30]. Tokio modelio algoritmas identifikuoja visus tekste esančius būdvardžius, priskiria jiems SO įvertį remiantis turimu semantiniu būdvardžių žodynu ir įprastai be jokių emocinių svorių susumuoja viso teksto įverčius.

Daiktavardžiai, veiksmažodžiai bei prieveiksmiai taip pat daro įtaką sakinio semantinei orientacijai [40]. Abu *1a* ir *1b* pavyzdžiai nusako tą patį įvykį, tačiau prieveiksmis pažymėtas pliuso ženklu *1a* sakinyje turi teigiamą poliariškumo emociją, kai *1b* sakinyje minuso ženklu pažymėtas prieveiksmis bei veiksmažodis nešasi neigiamą poliariškumo emociją.

1a Jis *laimingai*$^{+}$ pasakojo apie savo darbus.

1b Jis *išdidžiai*$^{-}$ *šūkavo*$^{-}$ apie savo darbus.

Pažymėtina, kad sentimentiniai žodynai yra sudaromi ne tik iš pavienių žodžių, bet ir iš nuoseklių žodžių junginių, kurie tik esant kartu išreiškia tam tikrą poliariškumą ir SO įvertį, pavyzdžiui, kaip *„baisiai gražus“* arba *„nesveikai geras“* [39].

1.2. Sentimentų žodynu pagrįsto metodo pranašumai

Anksčiau jau buvo apibendrinta, kad teksto sentimentinės analizės uždavinys gali būti sprendimas dvejais būdais. Visgi teksto klasifikacija pagrįstas metodas turi kelis esminius trūkumus, kurie nėra tinkami šiam tyrinamajam darbui.

Teksto klasifikavimo metodas pagrįstas sisteminiu mokymusi, tai yra kai algoritmui pateikiami dideli tam tikrų duomenų tekstų rinkiniai yra klasifikuojami pagal teigiamą ir neigiamą poliariškumą, naudojant unigramas, bigramas (angl. *n-gram*, *unigram*, *bigram*, tai iš žodžių porų sudarytas masyvas, skaidant analizuojamą sakinį, kur porą sudaro n šio sakinio žodžių, pavyzdžiui, unigrama: $n = 1$ nario poromis suskirstytas sakinys; bigrama: $n = 2$ narių poromis suskirstytas sakinys),

taip pat atsižvelgiant į kalbos dalis. Įdomu tai, kad pati sėkmingiausia strategija yra naudoti unigramas [28], išlaikant žodžio individualų SO įvertį. Nors ir teksto klasifikacija pagrįsti metodai teksto poliariškumą nustato su aukštu tikslumu, pakeitus analizuojamų tekstų tematiką, rezultatai ženkliai suprastėja [5]. Priežastis paprasta – teksto klasifikatoriai būna apmokyti tik vieno tipo tekstų analizei, o pakeitus jų tematiką bei auditoriją, įdiegtos algoritmo strategijos taps nebeveiksmingos.

Kitas teksto klasifikatoriaus neretai netinkamai įvertinamas aspektas – tai sakinį kuriantys kalbiniai elementai. Atliekant SO įvertį svarbu atsižvelgti į kontekstinius keitiklius, kurie modifikuoja raktinių žodžių SO įvertį. Pavyzdžiui, neiginiai, (pvz., *ne laimingas*), stiprinimo (pvz., *labai laimingas*), silpninimo žodeliai (pvz., *mažai laimingas*) [26]. Klasifikatorius, deja, negali įvertinti šių įvardytų atvejų, kol jam nebus pateikta pakankamai mokymosi pavyzdžių.

Spręsti visas šias minėtas problemas gali hibridiniai sentimentų analizės modeliai, naudojantys abu algoritmus. Vienas iš tokių modelių aprašomas Cardie ir Choi darbe [11], kur tekstą pirma apdoroja paleidžiamas klasifikatorius, įvertinantis sakinių poliariškumą, o tik antru etapu yra identifikuojami visi kontekstiniai keitikliai, tokie kaip inversijos, stiprinimai arba silpninimai.

Šiame darbe sukurto algoritmo veikimo principas yra pagrįstas sudarytu lietuvišku sentimentų žodynu, o taip pat kontekstinių keitiklių įvertinimo bei sakinių semantinės orientacijos įverčių susietumo modelių naudojimu. Visa tai atlieka galutinį teksto emocinio lygio įvertinimą.

1.3. Emocijos išreiškimas emotikonais

Prasidėjus WEB 2.0 erai, kai žmonės pradėjo vis dažniau naudotis el. paštu, išmaniaisiais telefonais bei socialinėmis platformomis, bendravimo formatą papildė grafiniai simboliai – emotikonai, (anl. *emojis*). Jais vartotojai gali išreikšti įvairius veiksmus (pvz., bėgti), objektus (pvz., medis), veido mimikas, netgi gestus (pvz., iškeltas nykštys). Jų populiarumas toks didelis, kad Prancūzijoje daugiau nei pusė visų internetinių žinučių/pranešimų parašytos naudojant emotikonus [12].

Ši auganti tendencija pritraukė tiek socialinių, tiek kompiuterių mokslų atstovus [23, 13]. Buvo rasta, kad emotikonų naudojimas ir jų pasirinkimas gali priklausyti nuo vartotojo demografinio profilio (kultūra, lytis, amžius) [19] bei komunikuojančių žmonių santykio tipo bei erdvės [21]. Daug dėmesio taip pat skirta galimam emotikonų panaudojimui kaip vienam iš sentimentinės analizės faktorių [24]. Dauguma tokių darbų savo duomenų bazes surenka iš socialinių tinklų, kadangi tokios platformos turi didelį vartotojų kiekį, kuris yra įvairialypis demografine, tautine ar pažiūrų įvairove. Taip pat šie įrašai lengvai prieinami (t.y. naudojant jau paruoštą pačių platformų programinę sąsają).

Iki šiol atlikti darbai koncentravosi į tokius elementus kaip paliktų komentarų skaičius, įrašų pasidalinimo ar „patinka“ (angl. *Like*) kiekį, tokiu būdu praturtinant SO analizę [4]. Taip pat kaip įrašuose ar jų komentaruose esantys emotikonai gali nustatyti to pačio įrašo SO įvertį. Kai kuriuose darbuose tai buvo atliekama priskiriant tam tikrą emocinį svorį visiems grafiniams simboliams ir taip juos sumuojant per visą tekstą ar komentarų masyvą [24]. Deja, toks būdas nėra universalus dėl vis naujai pristatomų emotikonų bei negalėjimo įvertinti situacijų, kai dėl tam tikrų įvykių pasaulyje, kai kurie grafiniai simboliai įgauna kitokią prasmę. Tačiau „Facebook“ pristačius papildomus 5 reakcijos emotikonus prie „patinka“ opcijos, sentimentinis uždavinys įgavo kur kas universalesnį analizės įrankį. Nuo šiol šios socialinės platformos vartotojai gali konstruktyviau išreikšti savo emocijas, tuo tarpu turinio kūrėjai – ar tai būtų įmonės, ar visuomenės veikėjai, ar tiesiog tie patys vartotojai – gali geriau suprasti savo auditoriją bei koreguoti savo pateikiamo turinio strategiją [10, 6]. Visa ši atsiradusi dinamika atvėrė naujas galimybes kuriant sentimentinės

analizės algoritmus bei juos optimizuojant [27]. Pavyzdžiui, buvo pastebėta, kad šis „Facebook“ 5 reakcijų papildymas gerai atspindi skaitytojų komentaruose naudojamus grafinius simbolius [33], o tai leidžia taupyti veikimo laiką, neanalizuojant komentarų sekcijos, bet norint suvokti bendrą įrašo kontekstą. Naujosios „Facebook“ reakcijos įgavo ir praktinį panaudojimą. 2016 metų JAV prezidento rinkimų metu politologai, analizuodami žmonių naudojamus emotikonus bei reakcijas į vykstančius debatus ar kandidatų interviu, sėkmingai galėjo nuspėti rinkimų eigą [32]. Biržos brokeriai taip pat stebėdami kurie emotikonai buvo naudojami ties akcijos kainą galimai veikiančia naujiena, koregavo savo pirkimo / paradavimo algoritmus [29, 16].

Šiame darbe „Facebook“ reakcijos buvo naudojamos tirti jų kuriamą sentimentinę informaciją, kuri vėliau buvo apdorota kuriant emotikonų SO modelius ir tokiu būdu gerinant *SentiII* algoritmo efektyvumą.

2. *SentiII*, SO įverčio nustatymo algoritmas

Sukurtas modelis / algoritmas *SentiII* buvo visiškai įgyvendintas Python 2.7 programavimo kalba naudojant tik tris pagalbines bibliotekas: matematinę biblioteką *numpy*, skirtą optimizuoti kodo veikimo trukmę dirbant su dideliais masyvais, ir teksto lyginimo bei tvarkymo bibliotekas ($difflib$, *re* atitinkamai), kurios atlieka suderinamumą tarp sentimentinio žodyno bei tekstą sudarančių žodžių, identifikuodamos SO nešančius sakinio narius. Kodo naudojimo instrukcijas bei vartotojo sąsajos paaiškinimą galima rasiti darbo priede (žr. priedą B). „Facebook“ emotikonų modelis taip pat buvo suprogramuotas naudojant Python 2.7 kalbą be papildomų bibliotekų (šio modelio veikimas yra aptartas vėlesniame skyriuje). *SentiII* algoritmas rėmėsi dvejomis prielaidomis:

- žodžio, frazės, sakinio ar viso teksto semantinė orientacija turi skaitinę išraišką. Šiame darbe buvo pasirinkta skalė nuo -3 (labai neigiamas emocijos įvertis) iki 3 (labai teigiamas emocijos įvertis);
- individualių žodžių bei frazių semantine orientacija daro įtaką bendram teksto kontekstui.

Verta paminėti, kad pamatiniu algoritmu nebuvo siekiama įgyvendinti pilnos kalbinės analizės, įtraukiant žodžių kontekstines reikšmes. Visgi įgyvendintas emotikonų modelių papildimas išsprendžia šią problemą, kadangi skaitytojas, kuris renkasi „Facebook“ reakciją, yra sąmoningas esamų įvykių bei konteksto atžvilgiu.
Sukurtas algoritmas remiasi dvejais sentimentų žodynais (žr. priedą D):

- sentimentų žodynas, kuriame pateikiami pagrindinės morfologinės formos žodžiai (tai būtų bendratis, vns. daiktavardis, vns. būdvardis ir t.t.) su atitinkamu SO įverčiu (nuo -3 iki 3);
- silpninančių / stiprinančių žodelių žodynas su jų atitinkamais silpninimo, bei stiprinimo koeficientais (nuo -1 iki 1).

SO įverčių pasirinktą rėžį (nuo -3 iki 3) galima paaiškinti taip:

- $SO = -3$ ir $SO = 3$ atitinkamai reprezentuoja visiškai neigiamą ir visiškai teigiamą SO emocinį įvertį. Pavyzdžiui, *nuostabus*$^{+3}$, *tragiškas*$^{-3}$;
- $SO = -2$ ir $SO = 2$ reprezentuoja neigiamą ir teigiamą SO emocinį įvertį. Pavyzdžiui, *taika*$^{+2}$, *melas*$^{-2}$;
- $SO = -1$ ir $SO = 1$ reprezentuoja tokius neigiamus ir teigiamus žodžius, kurių poliariškumas gali būti neaiškus arba ginčytinas. Pavyzdžiui, *pagirti*$^{+1}$, *problema*$^{-1}$;

- SO įverčiai taip pat gali būti tarp aukščiau išvardytų SO įverčių kaip racionalūs skaičiai, jei esantis žodis / frazė yra tarp dviejų apibrėžimų. Pavyzdžiui *medis*$^{+0.5}$ yra neutralaus poliariškumo žodis, tačiau lietuvių tautosakoje jis gali simbolizuoti stiprybės, gyvybės simbolį.

2.1. Sakinio analizės logika

Algoritmo veikimas pagrįstas 4 etapais. Pirmame etape (žr. 1 algoritmą) apdorojamas esamas tekstas, keičiant visas lietuviškas raides į jų lotyniškų raižių atitikmenis (t.y „ą“ į „a“, „ž“ į „z“ ir pan.), naikinami skyrybos ženklai išskyrus taškus (palaikant teksto struktūros informaciją), šauktukus, daugtaškius bei panašius ženklus, kurie gali suteikti SO informaciją. Vėliau tekstas skaidomas į sakinius laikant tašką kaip skiriamąjį elementą. Taip sudaromas sakinių masyvas, kuris pavadintas „$newtext$“. Algoritmas toliau dirba su kiekvienu sakiniu atskirai, vykdydamas atitinkamus etapus.

1 algoritmas. Pirmo etapo įgyvendinimo pseudokodas.

Įvestis: $text$: analizuojamas originalus tekstas; (lt_i, lot_j): lietuviškų bei atitinkamai lotyniškų raidžių keitinys; $(symb_i,' ')$: skyrybos ženklų į tuščius tarpus keitinys

Išvestis: $newtext$: modifikuotas tekstas

1: **for all** sakiniams sak **in** $text$ **do**
2: $newsak \leftarrow sak.replace(lt_i, lot_j)$
3: $newsak \leftarrow newsak.replace(symb_i,' ')$
4: $newtext_i \leftarrow newsak\ i \in [0, \texttt{len}(text)]$
5: **end for**
6: **return** $newtext$

Antrame etape, naudojant turimus žodynus, sakiniuose (apačioje nurodyta eiliškumo tvarka) identifikuojami:

- **Sentimentiniai žodžiai / frazės.** Tai daroma skaidant sakinį į n-gramas. Ar tai bus unigrama, ar bigrama ir t.t. priklausys nuo ieškomos žodyno frazės sudarančių žodžių skaičiaus. Pavyzdžiui, jei iš sentimentų žodyno masyvo paimtas narys yra „baisiai geras“, tai sakinys bus skaidomas į bigramas, kurios bus lyginamos su paimtu nariu. Kadangi tekste esantys žodžiai yra vartojami įvairiomis morfologinėmis formomis, o sentimentų žodynas buvo sudarytas tik su pagrindine forma (vns. daiktavardis, bendratis ir t.t.), algoritme buvo panaudota Python bibliotekos $difflib$ funkcija „SequenceMatcher“. Jos veikimas pagrįstas Levenšteino atstumo (angl. *Levenshtein distance*) skaičiavimu, tai yra kiek mažiausiai keitimų reikia atlikti, kad abu lyginami žodžiai visiškai sutaptų. Šios funkcijos išvestis yra koeficientas nurodantis sutapimo lygį η. Šiame darbe buvo pasirinkta, kad tekste ir sentimentų žodyne esančios frazės yra tos pačios, jei „SequenceMatcher“ funkcijos koeficientas yra daugiau nei 70%. Taigi sutapimo atveju frazė sakinyje pakeičiama simboliu: $''2\backslash z/''$, kur simbolio pradžioje esantis skaičius nurodo SO įvertį.
- **Silpninantys / stiprinantys žodeliai.** Kaip ir sentimentinių žodžių identifikavime, sakinys skaidomas į n-gramas priklausomai nuo silpninančių / stiprinančių žodelių žodyne esančios frazės ilgio. Žodyno masyvo narys tiesiogiai lyginamas su sakinį sudarančiomis pasirinkto dydžio n-gramomis. Radus atitikmenį, jis keičiamas simboliu: $0.5\backslash i/$, kur simbolio pradžioje esantis skaičius nurodo silpninimo / stiprinimo koeficientą, kuriuo bus veikiamas šalia esantis, bet ne toliau nei už 3 žodžių, identifikuotas sentimentinis žodis.

- **Inversijos.** Tikrinama, ar sakinys neturi žodelių arba priešdėlių „nebe“ arba „ne“, kurie pakeistų sentimentinio žodžio poliariškumą. Šiame darbe inversija buvo įvertinama paslenkant SO įvertį kodo pradžioje nurodytu poslinkio dydžiu (numatytas poslinkis $\delta = 2.5$). Rasta inversija buvo žymima „n“ raide ir įtraukiama į skaičiavimą tik tada, jei po inversijos ne daugiau kaip už dviejų neutralių žodžių (šis nustatymas gali būti keičiamas) esantys žodžiai / frazės buvo identifikuotos kaip: $''\backslash z/''$ arba $''\backslash i/''$. Ši „n“ raidė pridedama minėtų simbolių gale, pavyzdžiui, $''2\backslash z/n''$ arba $''2\backslash i/n''$.

Pirmą bei antrą etapus galima puikiai apibendrinti pateikus du pavyzdžius – originalų sakinį (žr. *Etapas 0*), bei tą patį sakinį atlikus antro etapo veiksmus (žr. *Etapas 2*).

Etapas 0 Tu labai nemėgsti ežerų, tačiau man patinka vanduo bei mėlyna spalva.

Etapas 1 Tu labai nemegsti ezeru taciau man patinka vanduo bei melyna spalva.

Etapas 2 Tu $''0.5\backslash i/''$ $''1\backslash z/n''$ ezeru taciau man $''1.5\backslash z/''$ $''1\backslash z/''$ bei $''1\backslash z/''$ spalva.

Turint simboliais struktūrizuotą sakinio formą atliekamas trečias etapas, kuriame skaičiuojamas SO įvertis, keičiant esamus simbolius į atitinkamas skaičiavimo formules (žr. 2a - 2f punktus).

2a $''x\backslash i/n'' \rightarrow *(1-x)*$, kur x – silpninimo \ stiprinimo koeficientas

2b $''x\backslash i/'' \rightarrow *(1+x)*$

2c $''x\backslash z/n''$, kur $x > 0 \rightarrow +(x-\delta)+$, kur δ – inversijos poslinkio dydis (kode numatytas kaip $\delta = 2.5$)

2d $''x\backslash z/n''$, kur $x < 0 \rightarrow +(x+\delta)+$

2e $''x\backslash z/'' \rightarrow +x+$

2f sakinio žodžiai, kurie liko neutralūs, yra ignoruojami.

Suskaičiavus gautas formules, rezultatas yra dalinamas iš identifikuotų sentimentų žodžių skaičiaus. Visų šių veiksmų įgyvendinimas atvaizduotas 2 algoritme.

Galima grįžti prie anksčiau nagrinėto pavyzdžio ir atvaizduoti trečiąjį etapą (žr. *Etapas 3*). Reikia atkreipti dėmesį, kad sakinys skaidomas į dvi skaičiavimo grupes. Šis dalinimas atliekamas algoritmo, kai identifikuotus sentimentinius elementus skiria daugiau nei du neutralūs žodžiai, tokiu būdu išlaikant sakinio emocinę struktūrą. Suskaičiavus trečiame etape gautas formules bei atsakymą padalinus iš rastų sentimentinių žodžių skaičiaus gaunamas šio sakinio $SO_{sak} \simeq 0.3$. Toks rezultatas reiškia, kad sakinys yra silpnai teigiamas / neutralus. Tai lėmė stiprų neigiamą pareiškimą sakinio pradžioje atsvėrę kitoje sakinio dalyje esantys trys teigiamo SO įverčio žodžiai.

Etapas 2 Tu $''0.5\backslash i/''$ $''1\backslash z/n''$ ezeru taciau man $''1.5\backslash z/''$ $''1\backslash z/''$ bei $''1\backslash z/''$ spalva

Etapas 3 $((1+0.5)\times(1-2.5))+(1.5+1+1) \simeq 0.3$

2 algoritmas. Trečio etapo įgyvendinimo pseudokodas.

Įvestis: sak: analizuojamas modifikuotas sakinys; $elem$: sakinio elementas / simbolis; mag: SO įvertis arba SO silpninimo / stiprinimo koeficientas

Išvestis: s_rez: SO įverčio skaičiavimo formulė

```
for all elem in sak do
    if '\i/n' in elem then
        s_rez ← s_rez + ' × (1 − mag) × '
    else if '\i/' in elem then
        s_rez ← s_rez + ' × (1 + mag) × '
    else if '\z/n' in elem then
        if mag > 0 then
            s_rez ← s_rez + ' + (mag − δ) + '
        else
            s_rez ← s_rez + ' + (mag + δ) + '
        end if
    else if '\z/' in elem then
        s_rez ← s_rez + ' + mag + '
    end if
end for
return s_rez
```

Išanalizavus praktinį atvejį bus apibrėžiamos galimos algoritmo skaičiavimo situacijos ir apibendrintos matematinėmis formulėmis:

$$\text{"}0.5\backslash i/\text{"}\ \text{"}1\backslash z/\text{"} \rightarrow (1 + N_{w_1}) \times SO_{w_2}\ ,$$

kur w_i yra sakinio i žodis, SO_{w_2} žodžio w_2 SO įvertis, kuris lygus 1, N_{w_1} stiprinimo daugiklis lygus 0.5;

$$\text{"}0.5\backslash i/\text{"}\ \text{"}1\backslash z/n\text{"} \rightarrow (1 + N_{w_1}) \times (SO_{w_2} - \delta)\ ,$$

kur δ yra inversijos poslinkis lygus 2.5, kadangi $S_{w_2} > 0$ prie δ atsiranda minusas;

$$\text{"}0.5\backslash i/n\text{"}\ \text{"}1\backslash z/\text{"} \rightarrow (1 - N_{w_1}) \times SO_{w_2}\ ,$$

kurioje matoma kaip algoritmas sprendžia inversijas, esančias prie silpninančių / stiprinančių žodelių, t.y. keičiant šių žodelių SO įverčio ženklą;

$$\text{"}1\backslash z/\text{"}\ \text{"}2\backslash z/\text{"}\ \rightarrow\ SO_{w_1} + SO_{w_2}\ ,$$

kur iš eilės einančių identifikuotų sentimentinių žodžių suminis SO įvertis yra jų atskirų SO įverčių suma.

Susumavus sakinyje identifikuotus sentimentinius elementus ir jų įverčius pagal taisykles, įvardytas viršuje, visa tai padalinama iš sakinyje identifikuotų sentimentinių žodžių skaičiaus:

$$SO_{sak} = \frac{\sum_{i,j}(SO_{w_i}, N_{w_j})}{\sum_{w_i \in sent.zodynas} i}\ ,$$

kur w_i, w_j identifikuoti žodžiai, atitinkamai turintys SO_{w_i}, N_{w_j} sentimentinį įvertį.

2.2. Sakinių susietumo logika

Įvertinus SO emocinį įvertį kiekvienam sakiniui tekste atskirai, būtų klaidinga išvesti SO vidurkį ir laikyti gautą rezultatą bendru viso teksto emocinio lygio įverčiu. Tokios strategijos pasirinkimas teigtų, kad šalia esančių poliariškai priešingų sakinių (pvz., $SO_i = -3$ ir $SO_{i+1} = 3$) bendras SO įvertis būtų 0, o tai reikštų neutralų tekstą. Taip pat visi teksto sakiniai būtų tarpusavyje sentimentiškai lygūs. Pavyzdžiui, pirmas sakinys veiktų šeštą sakinį taip pat kaip penktas sakinys šeštąjį. Šiame darbe ši problema buvo sprendžiama pasirenkant, kad kiekvienas prieš tai esantis sakinys nešasi tam tikrą susietumo SO svorį $\gamma(d)$, kur d – atstumas tarp analizuojamo ir praeities sakinio. Tokiu būdu, jei šiuo metu yra analizuojamas šeštasis sakinys, tai penkto sakinio SO įvertis darys didesnę įtaką einamojo SO įverčiui nei pirmo sakinio SO įvertis ir šis skirtumas bus aprašomas susietumo svoriu $\gamma(d)$. Ši strategija remiasi idėja, kad skaitantis žmogus labiausiai atsimena vėliausiai gautą informaciją.

Taigi algoritmui įvertinus pirmus du teksto sakinių SO įverčius pradedamas ketvirtas etapas – skaičiuojamas einamasis SO įvertis pagal pasirenkamą sakinių SO susietumo priklausomybę $\gamma(d)$. Galutinis teksto SO įvertis bus einamojo SO įverčio paskutinė vertė ir tai galima aprašyti formule (taip pat žr. į 3 pseudokodą):

$$SO_n = SO_{now} + \frac{\sum_{i=0}^{n-1}(SO_i \times \gamma(d))}{\sum_{i=0}^{n-1} i} \qquad d = (n-1) - i,$$

kur SO_{now} – analizuojamo sakinio SO įvertis, i – buvusių sakinių eilės numeris ir jų atitinkamas SO_i įvertis, SO_n – einamojo SO įvertis, $\gamma(d)$ – sakinių SO susietumo priklausomybė.

3 algoritmas. Ketvirto etapo įgyvendinimo pseudokodas.

Įvestis: d: atstumas tarp analizuojamo ir praeities sakinio; $SO_i[i \in 0..n-1]$: einamųjų SO įverčių masyvas be skaičiuojamo sakinio įverčio; $\gamma(d)$: sakinių susietumo svoris; SO_{now}: analizuojamo sakinio SO įvertis

Išvestis: SO_n: einamojo SO įvertis

1: $temp\_SO = 0$ kuris perskaičiuos einamąjį SO įvertį
2: **for all** SO_i **in** $SO_i[i \in 0..n-1]$ **do**
3: $\quad d \leftarrow (n-1) - i$
4: $\quad temp\_SO \leftarrow temp\_SO + SO_i \times \gamma(d)$
5: **end for**
6: $SO_n \leftarrow SO_{now} + temp\_SO/(n-1)$
7: $SO_i[n] \leftarrow SO_n$ papildomas enamasis SO masyvas n-tuoju, kątik rastu nariu
8: **return** SO_n

Šiame darbe buvo nagrinėtos 5 galimos sakinių susietumo svorių priklausomybės, kurios buvo ištirtos 3 skyriuje:

1 $\gamma(d) = \frac{1}{d}$ – tiesinė (t.y. SO įvertis atvirkščiai proporcingas praeitų sakinių atstumui);

2 $\gamma(d) = \frac{2}{d}$ – tiesinė su daugikliu;

3 $\gamma(d) = \frac{1}{d^2}$ – kvadratinė (SO įverčio priklausomybė mažėja kvadratu);

4 $\gamma(d) = \frac{1}{e^d}$ – eksponentinė (priklausomybė nyksta eksponentiškai);

5 $\gamma(d) = \frac{\sqrt{2}}{\sigma\sqrt{\pi}} \times \exp\left(-\frac{x^2}{2\sigma^2}\right)$ $x = d \times D(d)$ – pusinio normaliojo skirstinio (priklausomybė nyksta pagal tankio funkciją).

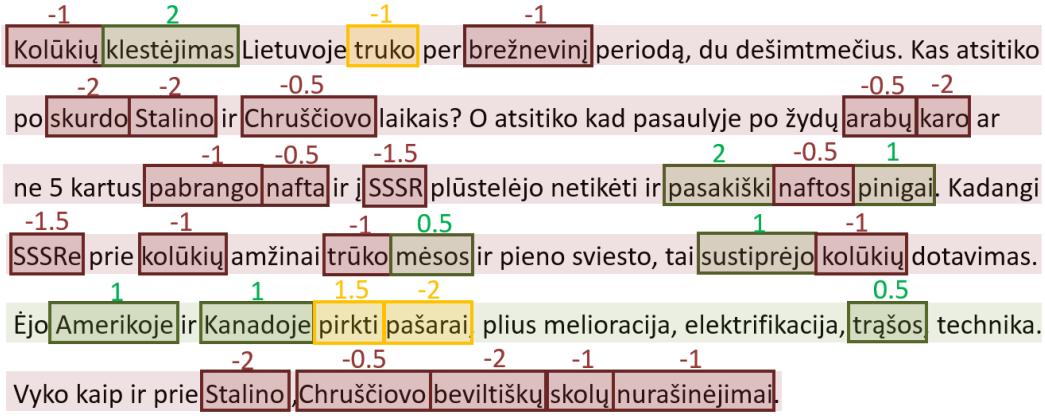
3. *SentiII* algoritmo testavimas

3.1. *SentiII* veikimo pavyzdys

Algoritmas buvo testuojamas naudojant įrašus iš socialinės platformos „Facebook“ puslapio „Laikykitės ten“ [1], kurio turinį sudaro aktualijos, dažniausiai pateikiamos satyros forma. Bus panagrinėtas algoritmo veikimas išanalizavus trumpą fragmentą, pateiktą 1 paveikslėlyje. Numatyti algoritmo nustatymai buvo:

1. Sakinių susietumo priklausomybė aprašoma pusiniu normaliuoju skirstiniu, kur $\sigma = 1$ (t.y. „*scale*“ parametras);
2. Toleruojamas dviejų žodžių frazių sutapimo koeficientas - $\eta = 70\%$;
3. Inversija paslenka SO įvertį $\delta = 2.5$ balais;
4. Inversija yra ignoruojama, jei atstumas tarp jos ir identifikuoto sentimentinio elemento daugiau kaip vienas neutralus žodis;
5. 1 žodis yra mažiausias neutralių žodžių skaičius tarp identifikuotų sentimentinių žodžių, kad sakinys būtų skaidomas į naują SO skaičiavimo grupę (žr. į 2.1 skyrelį).

1 iliustracijoje žalia spalva pažymėti sakiniai vaizduoja teigiamo poliariškumo sakinį, raudona – neigiamo poliariškumo. Apibrėžti žodžiai ar frazės yra kertiniai sentimentinę informaciją nešantys žodžiai, o skaičiai esantys virš šių žodžių rodo SO įvertį. Geltona spalva apibrėžti žodžiai yra neteisingai klasifikuoti elementai.



1 pav. Grafinis algoritmo veikimo reprezentavimas.

Gautos ištraukos $SO_n = -1.4$ identifikuoja tvirtą neigiamą poliariškumą. Visgi algoritmas suklasifikavo 3 žodžius neteisingai. Žodis *„truko“* tekste vartojimas kaip laiko trukmės įvardijimas,

tačiau algoritmas šį žodį suderino su žodžiu *trūkumas*. Tai įvyko dėl to, kad algoritmo pirmame etape lietuviškos raidės yra keičiamos į jų lotyniškus atitikmenis (taip supaprastinat darbą su netaisyklingais Lietuvių kalbos įrašais) su galimybe, kad žodžio reikšmė taip pat gali pakisti. Kiti du žodžiai *„pirkti“* ir *„pašarai“* algoritmo buvo palaikyti atitinkamai kaip *„patikti“* ir *„pragaras“*. Tai įvyko dėl to, kad frazių sutapimo toleruojamas koeficientas buvo $\eta > 70\%$, tačiau $\eta < 82\%$. Tai galima ištaisyti padidinant toleruojamą sutapimo koeficientą iki $\eta = 82\%$ arba kuriant atskirą neutralių žodžių žodyną. Visgi abu būdai atneštų nemažai netikslumų: per aukštas sutapimo koeficientas gali neatpažinti sudėtingos morfologinės formos žodžių, o neutralių žodžių žodynas gali surinkti ne neutralius elementus su panašia morfologine forma. Visgi, ištaisius šias klaidas, ištraukos poliariškumas nepasikeičia $SO_n^{'} = -1.3$.

3.2. Optimali *SentiII* konfigūracija

Algoritmo veikimo tikslumo (angl. *precision*), atkūrimo (angl. *recall*, t.y. jautrio / išbaigtumo matas) bei F_1 balo (angl. *F-messure*, t.y. tikslumo matas, įskaitantis abu atkūrimo ir tikslumo matus) įvertinimui, buvo panaudoti 150 skirtingų įrašų iš trijų „Facebook“ puslapių: „Laikykitės ten“, Vilniaus mero bei Andriaus Kubiliaus paskyrų. Puslapių autorių įrašai buvo analizuojami keletą kartų, keičiant pradinius algoritmo parametrus, tokiu būdų įvertinat optimalią konfigūraciją.

Įrašai buvo pasirinkti vadovaujantis šiais kriterijais:

- Tekstas neturi būti trumpesnis nei 3 sakiniai, tokiu būdu įvedant standartą tekstų apimčiai;
- Teksto turinys neturi būti reklaminio pobūdžio, tokiu būdu paliekant tekstus, kur autoriai dalijasi tik savo nuomone;
- Tekstas turi nemažiau nei 300 „Facebook“ reakcijos emotikonų, taip išfiltruojant populiarius.

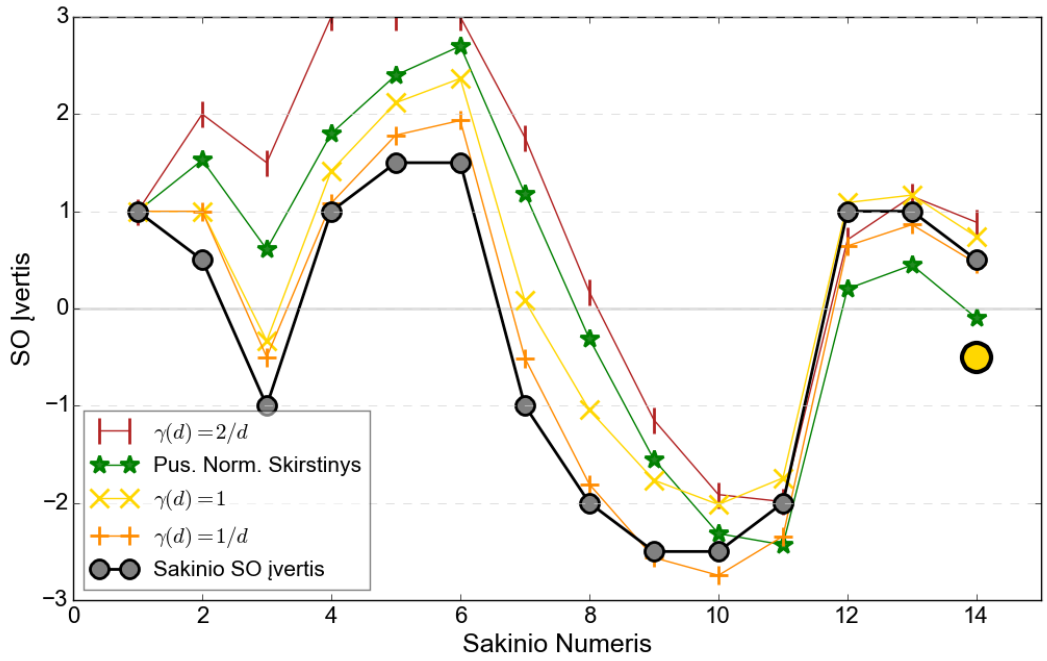
Visi šie 150 įrašų buvo peržiūrimi 5 nepriklausomų ekspertų (vienas iš jų buvo lietuvių kalbininkas), kurie kiekviename sakinyje turėjo identifikuoti galimus sentimentinius žodžius, kalbos keitiklius, inversijas bei pasakyti savo SO įvertį. Tuomet, keičiant SO susietumo priklausomybę $\gamma(d)$, toleruojamą sutapimo koeficientą η, algoritmas atliko savo SO analizę. Pirma bus palyginta, kaip algoritmas atpažino vertintojų rastus sentimentinius elementus, keičiant toleruojamą sutapimo koeficientą (koeficientas buvo keičiamas $\eta \in [0\%..100\%]$, 1% žingsniu). Tai buvo įvertinta skaičiuojant tikslumo, atkūrimo bei F_1 matus (žr. 1 lentelę). Matoma, kad aukščiausias F_1 balas pasiekiamas esant $\eta = 70\%$ toleruojamam koeficiento dydžiui. Tokį rezultatą galima paaiškinti aptarus koeficiento kraštines vertes. Esant mažam sutapimo toleravimo lygiui, algoritmas gali atlikti daug daugiau raidžių mainų, taip rasdamas ir negimininingus žodžius, atitinkamai atpažindamas visus tinkamus elementus (aukšta atkūrimo vertė, angl. *recall*), tačiau surinkdamas visus teksto žodžius (žemas tikslumas, angl. *precision*). Tai puikiai matosi, kai $\eta < 50\%$. Esant aukštai η vertei (pagal rezultatus $\eta > 86\%$) algoritmas atlieka minimalų raidžių keitimą, o tai mažina identifikuojamą populiaciją iki pagrindinę morfologinę formą turinčių sentimentinių žodžių. Tokiu būdu algoritmo tikslumas didėja, mažėjant atkūrimo vertei.

Pasirinkti tinkamą sakinių susietumo priklausomybę $\gamma(d)$ tolimesniam tyrimui buvo išrinkti 10 ilgiausių tekstų iš pradinių 150. Kiekvienam įrašui bei jį sudarantiems sakiniams vertintojai turėjo suteikti jų nuomone esamą SO įvertį. Tuomet kiekvienam įrašui buvo nubraižomi SO slenkančio įverčio grafikai naudojant skirtingus $\gamma(d)$, taip įvertinant susietumo priklausomybės trūkumus bei pranašumus.

1 lentelė. Algoritmo tikslumo, atkūrimo bei F_1 balo vidurkinės vertės, keičiant frazių sutapimo toleruojamą koeficientą.

Sutapimo toleruojamas koeficientas η, %	Tikslumas (angl. *precision*), %	Atkūrimas (angl. *recall*), %	F_1 balas (angl. *F-messure*)
30%	23%	100%	0.375
40%	25%	100%	0.401
50%	32%	97%	0.483
60%	36%	83%	0.503
70%	64%	90%	0.751
75%	59%	73%	0.657
80%	65%	67%	0.656
85%	72%	60%	0.655
90%	93%	47%	0.622
95%	100%	33%	0.499

Pateikta 2 iliustracija atvaizduoja minėtą SO slenkančio įverčio grafiką vienam įrašui iš 14 sakinių, kur abscisių ašyje įvardijamas sakinio eilės numeris, o ordinatės ašyje SO įvertis. Linija



2 pav. SO slenkančio įverčio grafikas, kuriame atvaizduojamas skirtingos $\gamma(d)$ strategijos. Iliustracijos dešinėje esantis didelis burbulas žymi vertintojų skirtą bendrą (t.y. visų sakinių) teksto SO įvertį.

su apskritimo formos žymekliu grafike vaizduoja vertintojų suteiktus SO įverčius sakiniams, pagal kuriuos keturios pasirinktos $\gamma(d)$ priklausomybės (linija su „x“ žymekliu – nėra jokios svorių susietumo priklausomybės, linija su pliuso žymekliu – tiesinė priklausomybė, linija su brūkšnio žymekliu – tiesinė su daugikliu bei linija su žvaigždutės žymekliu – pusinio normaliojo skirstinio) skaičiuoja einamąjį SO įvertį.

Matoma, kad $\gamma(d) = \frac{1}{d}$ priklausomybė (kreivė su pliuso žymekliu) mažai kuo skiriasi nuo originalių SO įverčių (kreivė su apskritimo žymekliu). Tai galima paaiškinti pavyzdžiu, kad pirmas teksto sakinys pridės vos 11% savo SO įverčio ketvirtojo sakinio SO įverčio skaičiavimui (tai yra dėl to, kad $\gamma(3) = \frac{1}{3} \simeq 33\%$, kurį padalinus iš sakinių skaičiaus 3, turėsime $\simeq 11\%$). Tuo tarpu $\gamma(d) = \frac{2}{d}$ priklausomybė (kreivė su brūkšnio žymekliu) elgiasi atvirkščiai – dėl koeficiento skaitiklyje (kuris $\neq 1$), įrašo pradžioje SO įvertis yra greitai prisotinamas teigiamu arba neigiamu poliariškumu. Esant ilgesniam tekstui, SO įvertis toliau išlieka šališkas pradžioje vyravusiam poliariškumui. Priklausomybė $\gamma(d) = 1$ buvo pasirinkta kaip palyginimas, kaip atrodytų SO įvertis teigiant, kad analizuojami sakiniai neturi svorinės „emocinės atminties“ perskaitytiems sakiniams, t.y. visi sakiniai vienodai svarbūs nepaisant to kaip senai jie buvo „skaityti“. Tokios logikos esminį trūkumą galima pamatyti 10, 12, 14 sakiniuose. Kadangi šiuo $\gamma(d)$ atveju sakinių susietumas nėra paremtas skirtingais svoriais, SO įverčio posūkio taškuose (t.y. 10, 12, 14 sakiniuose) einamasis SO įvertis supanašėja su pradiniais sakinių įverčiais, taip prarasdamas teksto „emocinę atmintį“ (dar vienas pavyzdys pateikiamas A priede esančioje 17 iliustracijoje).

Visų, kol kas įvardytų, $\gamma(d)$ kreivių paskutinis SO įvertis yra $1.5 - 2.0$ balais aukščiau nei buvo pateiktas vertintojų (žr. burbulą iliustracijos dešinėje), išskyrus pusinio normaliojo skirstinio (kreivė su žvaigždutės žymekliu). Ši strategija skiriasi tuo, kad svoriai yra normalizuojami pagal perlenkto-gauso pasiskirstymą. Šiame modelyje buvo naudoti $\sigma = 0$ (angl. *scale*) bei $\mu = 1$ (angl. *location*) parametrų vertės, su modifikacija, kad 1 sakinys atitinka dispersijos $D(x) = \sigma^2 \times (1 - \frac{2}{\pi})$ dydžio poslinkį pasiskirstymo abscisių ašyje. Galima pastebėti, kad šis variantas neturi minėtų problemų, kaip greitas SO įsisotinimas (žr. į 10, 11 numeriu pažymėtus sakinius) teksto pradžioje arba „emocinės atminties“ neturėjimas.

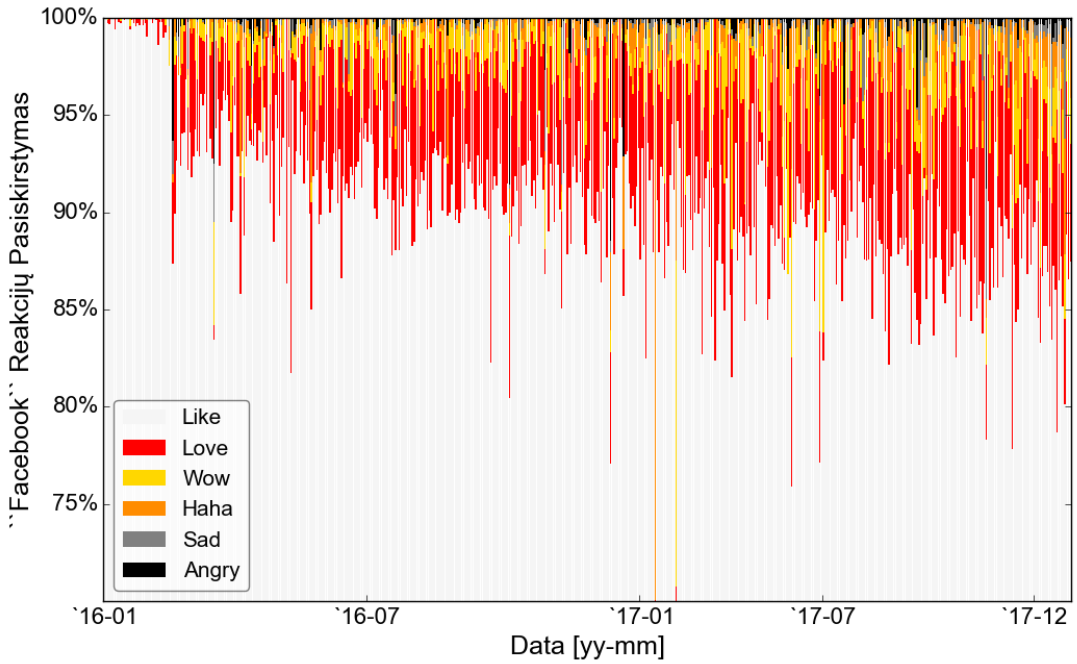
Pagal gautus rezultatus šiame skyriuje, tolimesnėje šio darbo eigoje *SentiII* algoritmas naudos šias konfigūracijas, kai sutapimo toleruojamas koeficientas lygus $\eta = 70\%$ bei sakinių susietumo pasirinkta strategija yra pusinis normalusis skirstinys $\gamma(d) = \frac{\sqrt{2}}{\sigma\sqrt{\pi}} \times \exp\big(-\frac{x^2}{2\sigma^2}\big)$, kur $x = d \times D(d)$.

4. Ar emotikonai teikia sentimentinę informaciją?

Šiuo metu „Facebook“ yra didžiausia socialinė platforma, kuri savo vartotojams yra pristačiusi reakcijos emotikonus, kuriuos galima naudoti emocinės būsenos, kylančios perskaičius publikuotus įrašus bei jų komentarus, išreiškimui. Pateikta iliustracija nr. 3 puikiai atvaizduoja šios grafinių simbolių augantį populiarumą viename iš populiariausių „Facebook“ puslapių „9GAG“, turintį net 38 milijoną sekėjų (dar vienas pavyzdys pateiktas A priede esančioje 18 iliustracijoje). Visgi kol kas emotikonai sudaro tik iki 30% visų naudojamų reakcijų. Tokį faktą galima paaiškinti tuo, kad jų pasirinkimo būdas yra neintuityvaus (t.y. palaikius pelytės žymeklį ties „patinka“ opcijos). Tačiau populiariose „Facebook“ grupėse minėti 30% gali būti $\simeq 3000$ individualių vartotojų reakcijų skaičius į vieną įrašą.

4.1. Surinkta duomenų bazė

Įvertinti kokią informaciją bei emocinius modelius gali suteikti 5 „Facebook“ reakcijos (t.y. be „Patinka“ opcijos), buvo surinkta virš 20000 įrašų iš 6 puslapių (2 naujienų agentūrų: „CNN“, „Foxnews“; 2 visuomenės veikėjų: „Bill Gates“ ir „Barack Obama“; 1 laisvalaikio: „9GAG“ bei

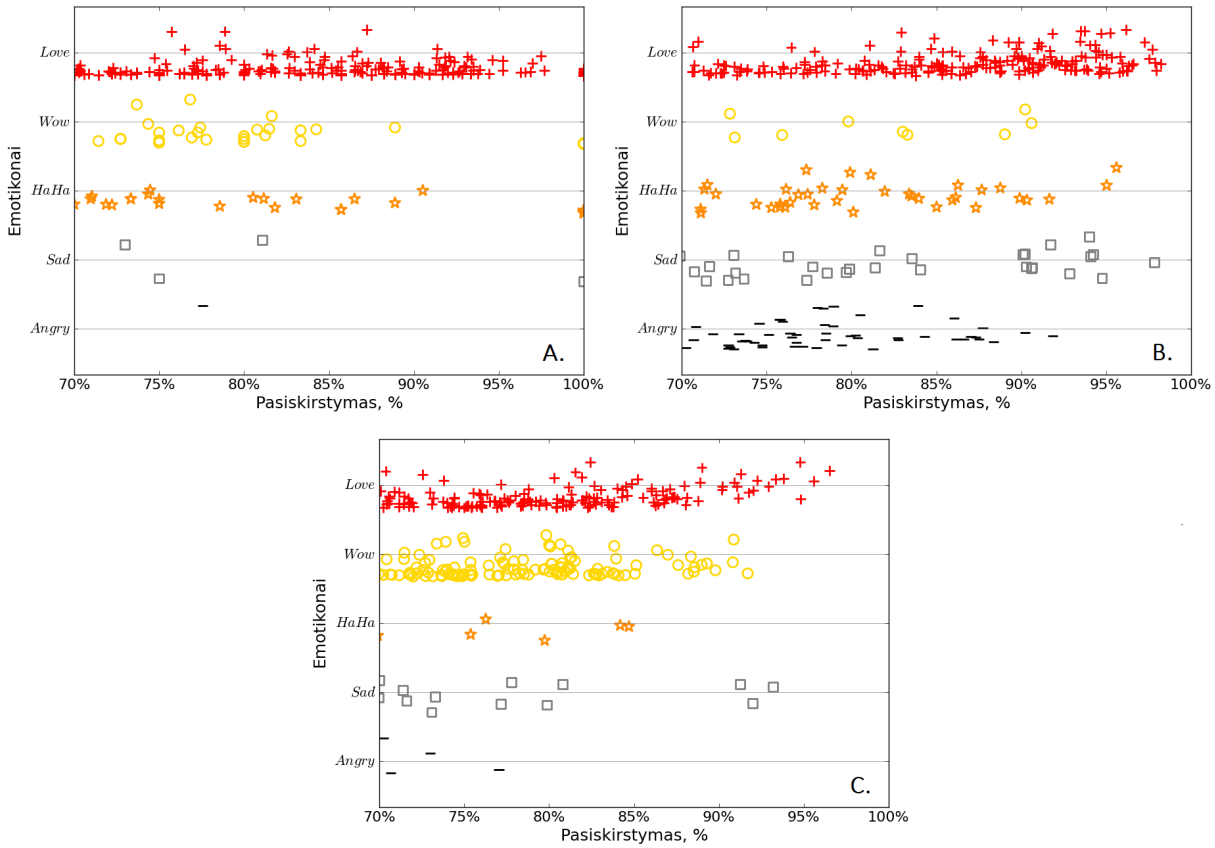


3 pav. „Facebook“ reakcijos emotikonų kiekio pasiskirstymas per dienos įrašus „Facebook“ „9GAG“ puslapyje nuo 2016 metų iki 2017 pabaigos.

gamtos aktualijų puslapių: „National Geographic“), kurie buvo publikuoti 2017 birželio – gruodžio mėnesių laiko periode. Kiekvienas puslapis turi virš kelių milijonų sekėjų. Visi šie įrašai turi virš 60 milijonų reakcijų, 7 milijonų komentarų bei buvo 13 milijonų kartų pasidalinti su kitais „Facebook“ vartotojais. Kiekvienas šis įrašas taip pat turi 11 duomenų segmentų: puslapio pavadinimą (kuriame įrašas publikuotas), patį tekstą, publikavimo laiką, kiek kartų įrašas buvo pasidalintas, komentarų skaičių ir atskirai 6 reakcijų kiekį. Šie visi duomenys buvo gauti suprogramavus Python kalbos skriptą, naudojantis „Facebook Graph 2.10“ programinę sąsają [2], kur išeities duomenys patiekiami *json* formatu. Kadangi visi šie įrašai yra pateikiami anglų kalba, jų SO įverčiui įvertinti buvo panaudota „Google natūralios kalbos“ [3] programinė sąsaja, kuriai pateikus įrašo tekstą kaip įvestį išvestyje gauname du kintamuosius – įvertį (angl. *score*) bei emocinį intensyvumą (angl. *magnitude*).

4.2. Emocinis įverčio nustatymas emotikonais

Norint įsitikinti, ar emotikonai iš tikrųjų atspindi įrašo turinio emocinę informaciją, buvo pasirinktos 3 „Facebook“ grupės, besiskiriančios savo įrašų turiniu: „9GAG“ (įvairūs humoro ir satyros tipo įrašai), „Foxnews“ (pasaulio naujienų ir aktualijų įrašai) bei „National Geographic“ (įvairių gamtos ir tiksliųjų mokslų naujienų bei sensacijų įrašai). Iš kiekvienos grupės buvo paimta 1000 naujausių jos įrašų analizei. Šiame tyrime buvo ignoruotas „patinka“ reakcijos buvimas, tokiu būdu 100% visų reakcijos emotikonų sudarė „WOW“, „Angry“, „Haha“, „Love“, „Sad“ bendra suma. Taigi, atvaizdavus minėtų trijų grupių emotikonų procentinį pasiskirstymą, rodant tiktai tą rėžį, kuriame esantys taškai atspindi reakcijas, sudarančias daugiau nei 70% visų įrašo reakcijų dalies, turimi visiškai skirtingi profiliai (žr. 4 iliustraciją). Reikėtų paminėti, kad 4 iliustracijoje atidėtų emotikonų taškų vieta ordinačių ašies atžvilgiu atspindi tos reakcijos kiekį, t.y. kuo aukščiau šis taškas yra, tuo daugiau tos reakcijos buvo kitų taškų tos pačios reakcijos atžvilgiu.



4 pav. „Facebook“ emotikonų kiekio pasiskirstymas per įrašą trijose skirtingose „Facebook“ grupėse, kur A. vaizduoja „9GAG“ grupės įrašus, B. – „Foxnews“ grupės įrašus ir C. – „National Geographic“ įrašus.

Pradėkime nuo „9GAG“ grupės emotikonų profilio (4 iliustracijos A dalis), kuriame „Love“ yra dominuojanti reakcija, o kitos, ypač „Sad“ ir „Angry“, turi vos kelis įrašus su didesne nei 70% visų reakcijų dalimi. Toks rezultatas visiškai atspindi šios grupės turinį. Tuo tarpu dienos aktualijomis besidalinanti naujienų grupė „Foxnews“ (žr. 4 iliustracijos B dalį) turi nemažai „Angry“ bei „Sad“ grupės emotikonų. Tokį faktą galima paaiškinti tuo, kad kiekvienas pasaulinis įvykis sulaukia įvairiausių nuomonių, o taip pat pačio naujienų portalo tikslas ir yra skelbti kontraversiškas sensacijas, tokiu būdu pritraukiant didesnę auditoriją bei jos reakciją. Šį teiginį galima paremti tuo, kad iš visų pavaizduotų grupių, ši turi daugiausiai įrašų su „Angry“ reakcijomis, sudarančių daugiau nei 70% visų įrašo emotikonų. Kitų reakcijų išsidėstymas atrodo panašiai. Likusios grupės „National Geographic“ emotikonų profilis, pavaizduota 4 iliustracijos C dalyje, išsiskiria nemažu savo „WOW“ reakcijos dažnumu. Didelė dalis šios grupės turinio yra naujausi moksliniai atradimai, sensacijos arba tiesiog įspūdingi gamtos vaizdai, o visi šie minėti dalykai atitinkamai reprezentuoja „WOW“ ir labai mažą „Angry“ ar „Sad“ emotikonų kiekį.

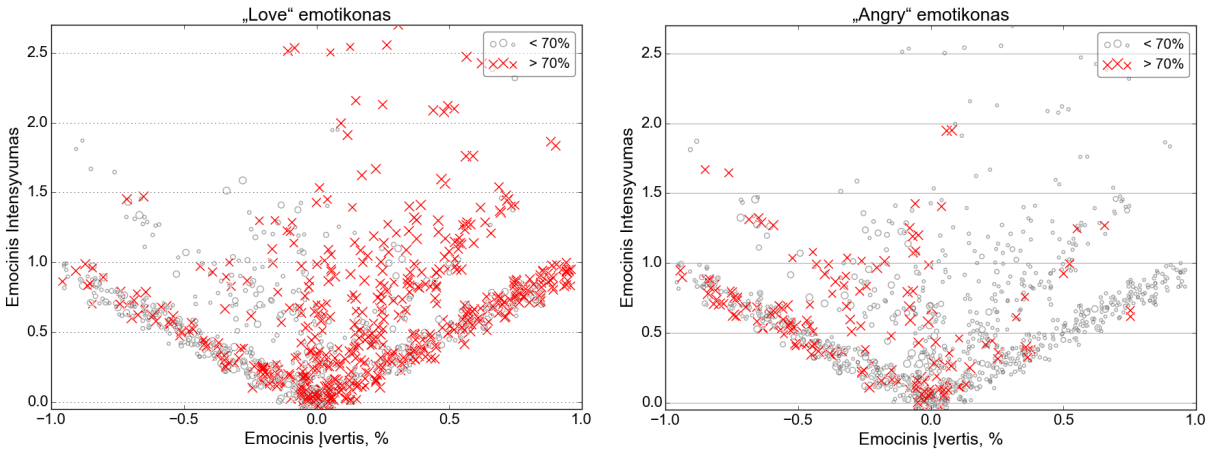
Įsitikinus, kad emotikonų profilio pasiskirstymas gali puikiai įvardyti puslapio publikuojamą turinį, atlikta šių puslapių turinio analizė įrašų lygyje, taip įvertinant, ar sentimentiniai modeliai gali būti apibūdinami emotikonų vartojimu. Šiam uždaviniui naudoti 6000 įrašų, kurie buvo pateikti „Google natūralios kalbos“ algoritmui SO įverčiams gauti. Kaip jau buvo minėta 4.1 skyriuje, „Google“ algoritmas turi kitokį emocinio įverčio pateikimo metodą: vartotojas gauna du parametrus. Pirma, tai pats įvertis (angl. *score*$\in [-1, 1]$), kuris atspindi teksto poliariškumą. Antra, tai emocinį intensyvumą apibūdinantis parametras (angl. *magnitude*$\in [0, +\infty]$), kuris yra proporcingas teksto ilgiui – tiek neigiama, tiek teigiama emocija didina šio parametro vertę. Geresniam šių

dviejų parametrų kombinacijų supratimui ir kaip juos teisingai interpretuoti pateikta sentimentų lentelė nr. 2.

2 lentelė. „Google natūralios kalbos“ sentimentų algoritmo rezultatų interpretavimas.

SO Įvertis	Įvertis (*score*)	Emocinis intensyvumas (*magnitude*)
Labai Neigiamas	-0.8	3.5
Labai Teigiamas	0.8	4.0
Silpnai Neigiamas	-0.8	0.5
Neutralus	-0.1	0.0
Mišrus	0.1	4.0

Gavus visų 5000 tekstų SO įverčių bei intensyvumo parametrus ir surinkus „Facebook“ 5 reakcijų kiekius tiems patiems tekstams, buvo nubraižyti emotikonų procentiniai pasiskirstymai SO matricoje. Visa tai yra grafiškai atvaizduojama 5 iliustracijoje, kur abscisių ašis nusako emocinį įvertį (*score*), o ordinačių – emocinį intensyvumą (angl. *magnitude*) (visa tai bendrai vadinsime SO įverčio matrica). Abu iliustracijoje esantys grafikai nusako konkretaus emotikono skirstinį, pateiktu atveju „Love“ kairėje ir „Angry“ dešinėje paveiksliuko pusėje. Šiuose grafikuose atidėti taškai, kurie reprezentuoja vieną unikalų įrašą, skiriasi dydžiu, kuris atspindi vaizduojamo emotikono procentinę dalį viename įraše. Kuo didesnis taškas, tuo didesnę procentinę dalį vaizduojama reakcija sudarė visų to įrašo reakcijų atžvilgiu. Kai vaizduojamas emotikonas sudaro daugiau nei 70% įrašo reakcijų, taško forma pakeičiama iš apskritimo į pliusą, tokiu būdu supaprastinat vizualinį reakcijų pasiskirstymą SO matricoje. Pirma, aptarkime kairiąją iliustracijos pusę – „Love“



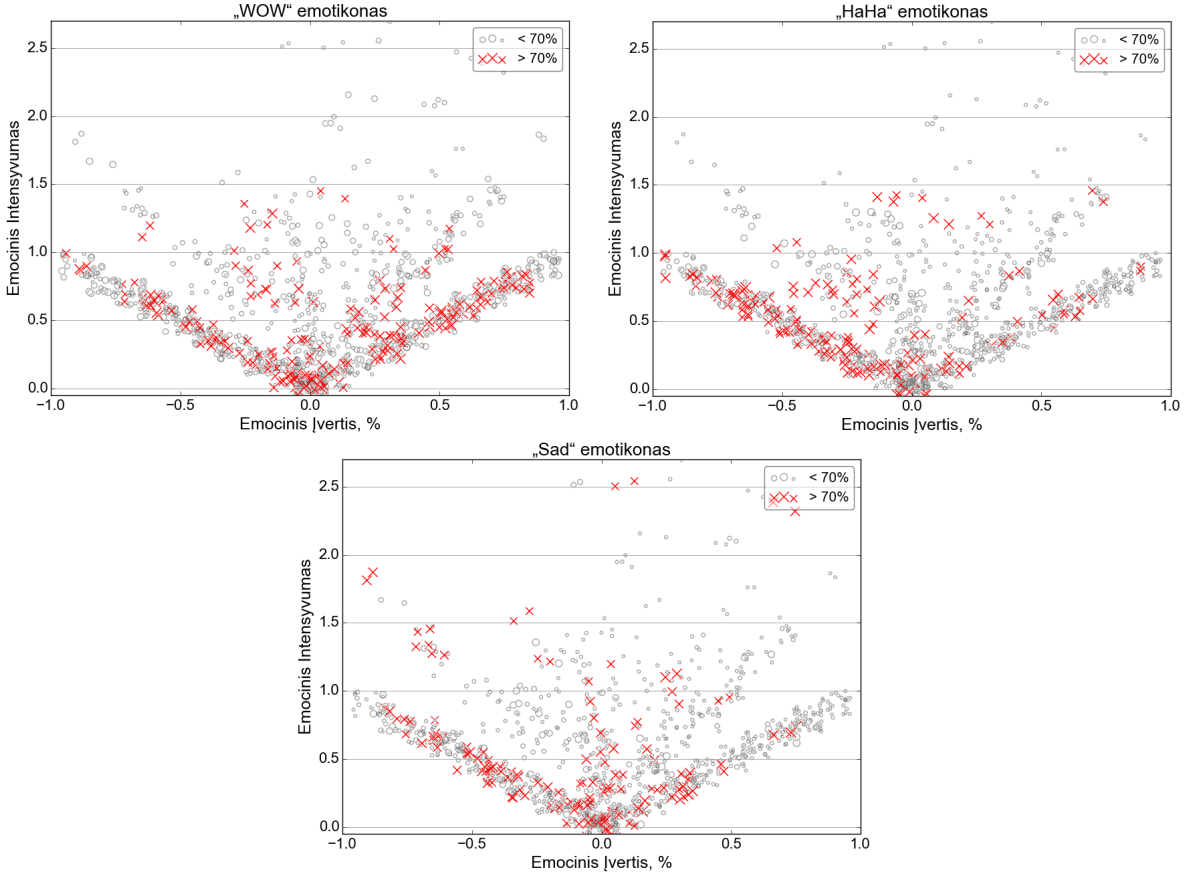
5 pav. „Love“ ir „Angry“ emotikonų „Facebook“ įrašuose procentinis pasiskirstymas SO matricoje, kuri buvo įvertinta naudojant „Google natūralios kalbos“ programinę sąsają.

skirstinį. Aiškiai matoma, kad beveik visi įrašai, kuriuose „Love“ sudaro $> 70\%$ visų reakcijų, išsidėstę teigiamo įverčio pusėje. Pliuso formos taškai, kurie pateko į neigiamą dalį, yra neteisingai suklasifikuoti tekstai, kurių dauguma yra motyvacinių kalbų tekstų tipo, kurie dėl savo struktūros turi neigiamo poliariškumo skiltį. Dešinėje pusėje matomas atvirkščias pasiskirstymas. Čia įrašai,

kur „Angry“ emotikonai sudaro $> 70\%$ visų reakcijų, pasislinkę į neigiamo įverčio pusę. Visus šiuos rezultatus galima logiškai pagrįsti, kadangi abi reakcijos žymi visiškai skirtingo poliariškumo emocijas. Tokiu būdu galima teigti, kad jei įrašo didžiąją reakcijų dalį sudaro vienas iš „Love“ arba „Angry“ emotikonų, SO įvertis atitinkamai turės būti teigiamas arba neigiamas.

Bus aptarta dar viena svarbi SO matricos grafiko vieta. Pastebima, kad abi reakcijos 5 iliustracijoje turi didelę pliuso formos žymeklių sankaupą ties *score*$\simeq 0.0$ rėžiu. Panagrinėjus detaliau „Google natūralios kalbos“ veikimo specifikas paaiškėjo, kad algoritmas neturi jokios sakinių susietumo priklausomybės arba $\gamma(d) = 1$. Tokiu būdu, jei tekstą sudarytų du sakiniai atitinkamai $SO_{sak1} = -1$ ir $SO_{sak2} = 1$, bendras teksto $SO_{tekstas}$ būtų lygus $SO_{tekstas} = \frac{-1+1}{2} = 0.0$, tačiau teksto emocinė vertė būtų didesnė nei nulis. Šis fenomenas taip pat matomas 5 iliustracijoje. Tai, kad dauguma tokios grupės įrašų yra $0 <$*magitude*< 0.5 rėžyje, parodo, kad šie įrašai yra iki kelių sakinių ilgio.

Atvaizdavus SO įverčių matricas likusiems „WOW“, „Sad“ bei „Haha“ emotikonams, nematomas griežtas išsidėstymas abscisių ašies atžvilgiu (žr. 6 iliustraciją). Galima pastebėti taip pat jau



6 pav. „WOW“, „Haha“ ir „Sad“ emotikonų „Facebook“ įrašuose procentinis pasiskirstymas SO matricoje.

aptartą fenomeną ties *score*$\simeq 0.0$ rėžiu. Panagrinėjus kiekvieną įrašą, kur tiriamasis emotikonas sudaro $> 70\%$ visų reakcijų, buvo atrastos įdomios tendencijos.

Pirmiausia bus aptarta „Facebook“ emotikonas „WOW“. Pastarasis yra naudojamas tiek prie teigiamo, tiek prie neigiamo poliariškumo įrašų. Vienintelė bendra savybė tarp šių priešingo poliariškumo įverčių yra ta, kad šiais tekstais dažniausiai perteikiama sensacija, netikėta žinia ar įvykis. Tai kelia nuostabą ir taip atitinkamai vartotojo emotikono pasirinkimą. Tokio pobūdžio

tekstai dažniausiai yra trumpi, su aiškia žinute (turintys keletą sentimentinių žodžių, o tai galima pastebėti ir iš mažų emocinio intensyvumo verčių (angl. $0 <$*magitude*< 1). Tuo tarpu „Haha“ ir „Sad“ emocijos yra pasislinkusios ties neigiamu emociniu įverčiu.

3 lentelė. Tekstų pavyzdžiai, kur emocinis įvertis (*score*)< -0.5 bei dominuojanti reakcija yra „Haha“.

Orginalus Sakinys	LT Vertimas	Emocinis įvertis
Detroit police officers accidentally go after undercover cops in drug bust gone wrong	Detroito policija atsitiktinai užpuolė užsimaskavusius policininkus, taip sugadindami operaciją	-0.5
"In the span of ten days, she completely nuked her reputation" Michael Hopkins said	Michael Hopkins teigimu „per 10 dienų ji visiškai sunaikino savo reputaciją“	-0.7
George Takei is now being accused of sexually assaulting a man who has already died	George Takei buvo apkaltintas seksualiai priekabiavęs prie žmogaus, kuris jau buvo miręs	-0.8

4 lentelė. Tekstų pavyzdžiai, kur emocinis įvertis (*score*)< -0.5 bei dominuojanti reakcija yra „Sad“.

Orginalus Sakinys	LT Vertimas	Emocinis įvertis
More than 300 people have died and at least 4000 people have been injured after a 7.3 magnitude earthquake	Daugiau nei 300 žmonių mirė ir bent 4000 buvo sužeisti per 7.3 balų žemės drebėjimą	-0.7
Hillerman was best known for his emmy award-winning work in the detective series, he died being 84 years old	Hillerman buvo gerai žinomas už jo Emmy apdovanojamais įvertintus detektyvo žanro serialus, jis mirė būdamas 84	0.4
Sam Dryden died this morning. He was curious about everyone, he loved everyone and so many people loved him back	Sam Dryden mirė šį rytą. Jis buvo smalsus, mylėjo visus, taip ir visi mylėjo jį	0.5

Norint paaiškinti šį skirstinį, buvo pateiktos dvi lentelės su tekstų pavyzdžiais, kur 3 lentelėje

rodomi įrašai, kai „Haha“ sudaro daugiau nei 70% visų reakcijų su *score*< -0.5, o 4 lentelėje pateikiami „Sad“ reakcijos tekstai su tų pačių verčių parametrais kaip „Haha“. Iš pateiktų pavyzdžių 3 lentelėje galima pastebėti, kad įrašai su „Haha“ emotikonais yra ironiško / sarkastiško turinio, kur žodžiai, stiprinantys šį sarkazmą, yra neigiamo sentimentinio įverčio nešėjai. Panagrinėjus įrašus iš teigiamos įverčio (angl. *score*) srities, pastebėta tokia pati tendencija, tik šiuo atveju teigiamo SO įverčio žodeliai stiptina teksto sarkazmą. Tai, kad dauguma „Haha“ reakciją turinčių įrašų atsirado neigiamoje poliariškumo pusėje, lėmė pasirinktos „Facebook“ grupės, kurios skirtingai perteikia informaciją, šiuo atveju ironiją, naudojant teigiamo arba neigiamo SO įvertį turinčius elementus. Galima teigti, kad aptartos „Haha“ reakcijos emotikonas gali būti puikiai pritaikytas ironijos atvejų atpažinimui.

Tuo tarpu „Sad“ reakcija identifikuoja kitą reiškinį – mirties ar netekties turinio susijusį kontekstą. Tai puikiai atsispindi 6 iliustracijos apatinėje SO matricoje, kur didžiąją dalį užimantys „Sad“ emotikonai išsidėstę neigiamo įverčio < 0 srityje (žr. į pirmą 4 lentelėje pateiktą pavyzdį). Visgi dalis pastebima ir teigiamoje *score* srityje, o tai būtų galima paaiškinti tik peržiūrėjus tokius įrašus. Du tokie pavyzdžiai yra pateikti 4 lentelėje. Galima pastebėti, kad nors ir įrašo kontekstas yra susijęs su liūdesiu bei mirtimi, likusio teksto dalyje dėstomi prisiminimai yra teigiamą SO įvertį turinčių žodžių rinkinys. Visa tai algoritmas vertina kaip teigiamą sentimentinį įvertį, nors įrašo potekstė yra netektis.

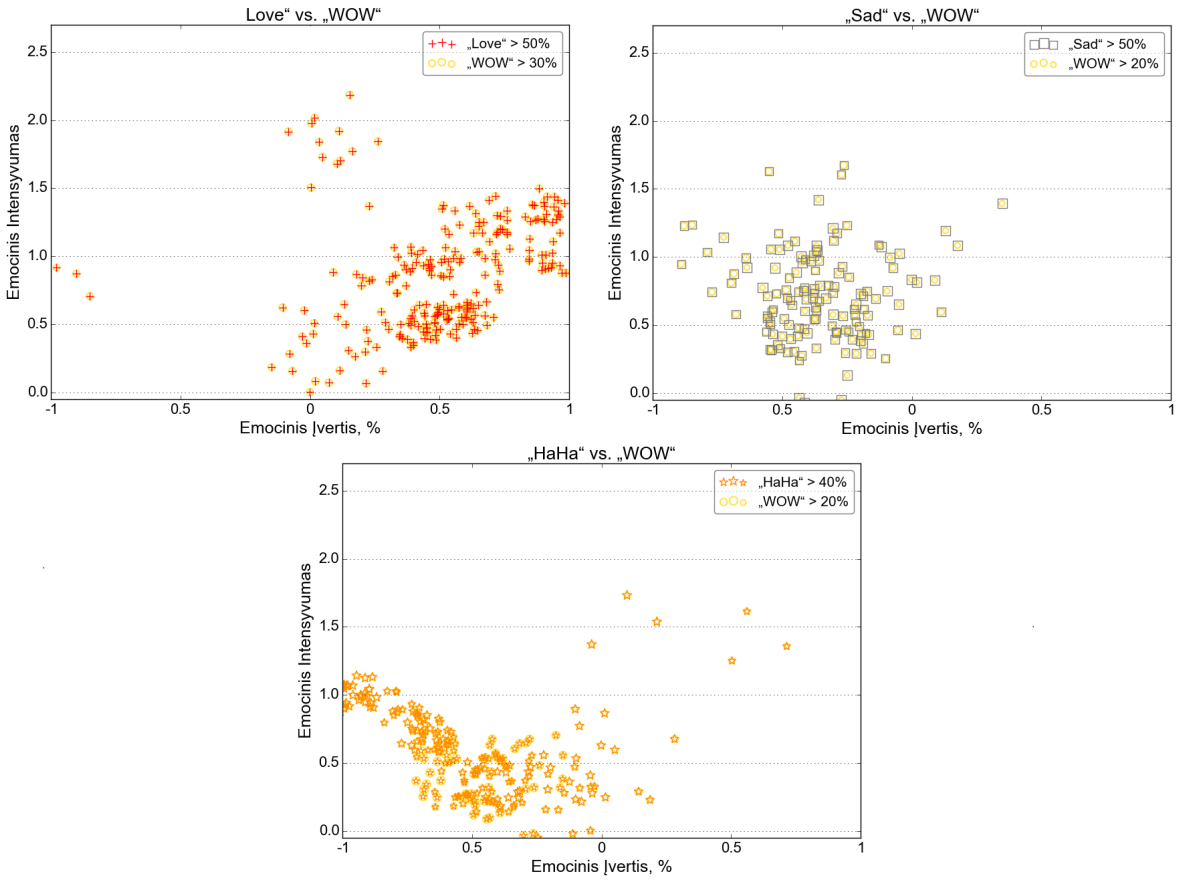
4.3. Emotikonų modeliai

Atlikus tyrimą kaip vartotojai naudoja „Facebook“ reakcijas bei kokią informaciją jų pasiskirstymas SO įverčių matricoje gali suteikti sentimentinei teksto analizei, rasta tiek „Google“ algoritmo trūkumų, tiek sprendimo būdų. Praeitame poskyryje įsitikinta, kad emotikonai turi perteikti teksto nešamą emociją ir tai iš tiesų daro. Kadangi emotikono pasirinkimas yra žmogaus atliekamas sprendimas, jis bus pagrįstas bei veikiamas dienos aktualijų, tekste esančių įvairių meninių raiškos priemonių bei istorinių ar praeities žinių. Minėti aspektai yra didžiausi sentimentinio uždavinio iššūkiai, kuriuos galima apibrėžti emotikonų modeliais. 4.2 skyriuje įsitikinta, kad:

- „Love“, „Angry“ bei „Sad“ reakcijos gan aiškiai nusako bendrą teksto poliariškumą;
- „Sad“ emotikoną turinčių tekstų turinys dažniausia yra susijęs su mirtimi;
- „WOW“ ir „Haha“ emotikonai gali būti naudojami tiek teigiamo, tiek neigiamo SO įvertį turinčiuose tekstuose;
- „WOW“ emocija dažniausiai randama, kuomet žmonės vertina sensacinius įrašus;
- „Haha“ emocija dažniausiai randama sarkastišku / ironiško turinio įrašų vertinime.

Visgi nei vienas iš išvardytų punktų negali nusakyti tiksliai nei SO įverčio stiprumo lygio, nei aiškaus teksto poliariškumo. Norėdami sukonkretinti emotikonų nešamą SO informaciją, turime atsižvelgti ir į kitų emotikonų dalį nagrinėjamame įraše. Peržiūrėjus emotikonų kiekių pasiskirstymą įrašuose pastebėta, kad (be „Patinka“ opcijos) vidutiniškai viena reakcija turi $\simeq 60\%$ visų emotikonų dalies. Jei mes sumuojame dvi didžiausią dalį užimančias reakcijas, gauname net $\simeq 92\%$ visų emotikonų dalies. Taigi, šiame tiriamajame darbe dirbta su dvejų reakcijų kombinacijomis. Panagrinėkime tris kombinacijas: „WOW“ – „Love“, „Sad“ – „WOW“ bei „Haha“ – „WOW“, pateiktas 7 iliustracijoje, kur abscisių ašis nusako emocinį įvertį (angl. *score*), o ordinačių – emocinį intensyvumą (angl. *magnitude*). Ant ašių esantys dydžiai buvo įvertinti naudojant „Google natūralios kalbos“ programine sąsaja. Atidėti taškai – tai unikalūs įrašai, kurių visų emotikonų populiaciją sudaro legendoje nurodyti du emotikonai.

Pavyzdžiui, apatiniame 7 iliustracijos grafike pavaizduoti įrašai, kurių „WOW“ reakcija sudaro daugiau nei 20% bei „Haha“ daugiau nei 40% visų įrašo reakcijų. Visų pirma pastebima, kad esant



7 pav. „WOW“ – „Love“, „Sad“ – „WOW“ bei „Haha“ – „WOW“ emotikonų kombinacijų pasiskirstymas SO matricoje.

papildomai reakcijai, „WOW“ įgauna konkretų poliariškumą. Kaip jau įsitikinta iš praeitų skyrių, „Love“ ir „Sad“ atitinkamai nurodė teksto poliariškumą, nepaisant kokia kita reakcija yra didžiausia pagal dalį (šiuo atveju esant „WOW“). Tuo tarpu „Haha“ su „WOW“ kombinacija turintys įrašai pasiskirstė $score < 0$ srityje. Šiuos pasiskirstymus galima paaiškinti peržiūrėjus tekstus, turinčius nagrinėtų emotikonų kombinacijų rinkinius. Apibendrinus 7 iliustracijoje pateiktus grafikus:

- „WOW“ – „Love“ – tai stipriai pozityvūs tekstais, pasakojantys apie įspūdžius arba teigiamą sensaciją;
- „Sad“ – „WOW“ tai stipriai negatyvūs tekstai, pasakojantys apie sukrečiančius įvykius;
- „Haha“ – „WOW“ tai negatyvūs tekstai ironizuojantys naujausius įvykius ar individų poelgius.

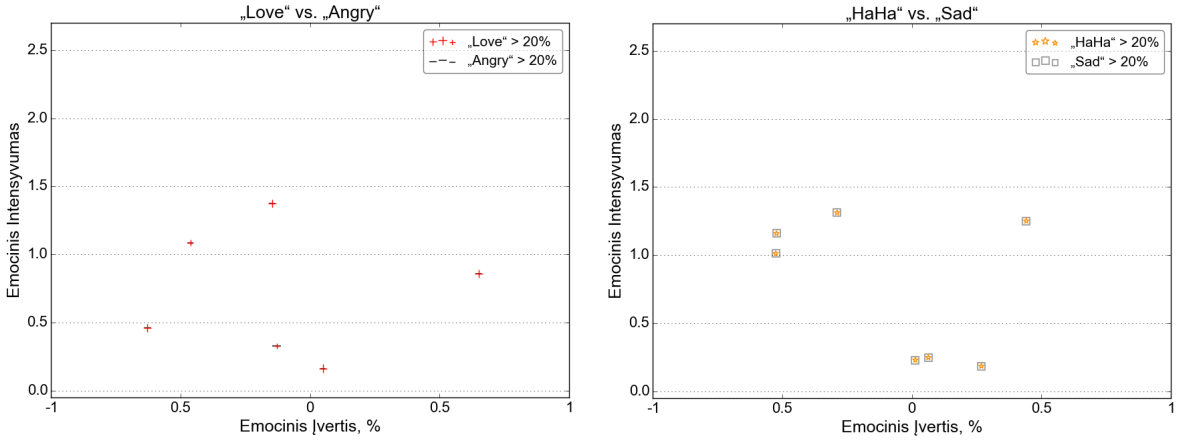
Ta pati analizė atlikta visoms kitoms emotikonų kombinacijoms. Rezultatai pateikti 5 lentelėje. Čia stulpelis „Emotikonų kombinacija“ nurodo aptariamą reakcijų derinį, „Poliariškumas“ ar derinys nusako teigiamą ar neigiamą poliariškumą (t.y. teigiamas ir neigiamas), „SO rėžis“ nurodo, kuriame emociniame įverčių ruože derinys egzistuoja (t.y. $\in [-1, 1]$), „Tekstų turinys“ apibendrina šių įrašų kontekstą. „Poliariškumo“ stulpelio vertė buvo įvertinta pagal tai, kurioje *scale* dalyje egzistuoja 75% visų įrašų, kai „SO rėžis“ buvo įvertintas pagal tai, kur egzistuoja 75% visų įrašų, esančių „Poliariškumo“ stulpelyje nurodytoje *scale* dalyje.

5 lentelė. Emotikonų kombinacijų suteikiama sentimentinė informaciją ir jų modeliai.

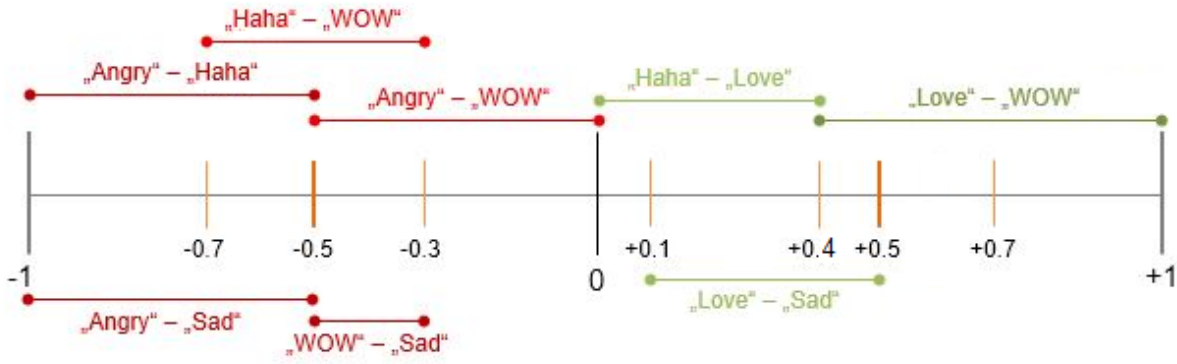
Emotikonų kombinacija	Poliariškumas	SO rėžis	Tekstų turinys
„Angry“ – „Haha“	Neigiamas	$[-1\ ,\ -0.5]$	Ironija, sarkazmas
„Angry“ – „Love“	–	–	Toks derinys neegzistuoja
„Angry“ – „WOW“	Neigiamas	$[-0.5\ ,\ 0.0]$	Neigiama sensacija
„Angry“ – „Sad“	Neigiamas	$[-1\ ,\ -0.5]$	Mirtis, masinė nelaimė, nelygybė
„Haha“ – „Love“	Teigiamas	$[0.0\ ,\ 0.4]$	Juokingas įvykis
„Haha“ – „WOW“	Neigiamas	$[-0.7\ ,\ -0.3]$	Nusivylimas įvykiu, individu
„Haha“ – „Sad“	–	–	Toks derinys neegzistuoja
„Love“ – „WOW“	Teigiamas	$[0.4\ ,\ 1]$	Teigiama sensacija, įspūdžių pasidalinimas
„Love“ – „Sad“	Teigiamas	$[0.1\ ,\ 0.5]$	Liūdna istorija su laiminga pabaiga
„WOW“ – „Sad“	Neigiamas	$[-0.5\ ,\ -0.3]$	Sukrečiantys įvykiai

Pastebima, kad kai kurie deriniai yra pažymėti kaip neegzistuojantys. Tai yra dėl to, kad iš visų ištirtų įrašų mažiau nei 1% turėjo tokią kombinaciją, kurioje abu emotikonai atskirai sudarytų daugiau nei 20%. Pavyzdžiui, „Angry“ – „Love“ derinys, kurio reakcijos visą laiką atspindi vieną iš priešingų emocinių poliariškumų, logiškai negali atspindėti to paties įrašo. Tai ir matome 8 iliustracijoje, kurioje vos keli įrašai turi tokių reakcijų derinį, o jų pasiskirstymas nėra tendencingas.

Įvertinus visas galimas 10 emotikonų porų ir aprašius radinius 5 lentelėje, taip pat pateikti „SO rėžio“ bei „Poliariškumo“ stulpeliai. Kiekvienai emotikonų kombinacijai šie stulpeliai nurodo sentimentinį poliariškumą bei apytikslį SO įvertį. Atvaizdavus kiekvieną šią kombinaciją pagal „SO rėžio“ vertes gaunamas 9 grafikas, kuriame visos 8 poros pilnai padengia SO rėžį (dviejų likusių porų „Love“ – „Angry“, „Haha“ – „Sad“ kombinacijos neegzistuoja), o vietose ir persidengia. Tokiu būdu šiame tyrime „SO rėžio“ bei „Poliariškumo“ stulpelių vertės bus naudojamos kaip emotikonų sentimentiniai modeliai, kureis koreguojami arba papildomi *SentiII* algoritmo gaunami SO įverčiai.



8 pav. „Love“ – „Angry“, „Haha“ – „Sad“ emotikonų kombinacijų pasiskirstymas SO matricoje.



9 pav. Emotikonų porų pasiskirstymas SO įverčio $\in [-1..1]$ rėžyje.

Norint pilnai paruošti emotikonų modelius, reikia išspręsti esminį skirtumą tarp „Google natūralios kalbos“ bei *SentiII* algoritmų teikiamų SO įverčio rezultatų. *SentiII* algoritmo rezultatas yra pagrįstas vienu skaičiumi, kuriuo emocijos stiprumas nusakomas skale $[0.. \pm 3]$ (su viršutine riba 3), o poliariškumas – teigiamu arba neigiamu ženklu prie skaičiaus. Tuo tarpu „Google“ algoritmo išvestis susideda iš dvejų parametrų, kur vienas nurodo poliariškumą (angl. *score*), o antras emocijos intensyvumą (angl. *magnitude*). Verta pastebėti, kad per visą tyrimą net $\simeq 99\%$ analizuotų įrašų pateko į $[0..2]$ emocinio intensyvumo rėžį, tad $magnitude_{virs} = 2$ laikysime viršutine riba. Taigi, naudojant emocinį įvertį bei intensyvumo viršutinę ribą kaip daugiklius, sprendimą galima pateikti šia formule:

$$SO_{SentiII} = score_{Google} \times magnitude_{Google} \times \frac{SO_{SentiIIvirs}}{magnitude_{virs}},$$

kur $SO_{SentiII}$, *SentiII* algoritmo SO įvertis, *score* bei *magnitude* „Google“ algoritmo SO įvertį sudarantys parametrai, $magnitude_{virs} = 2$ ir $SO_{SentiIIvirs} = 3$ „Google“ ir atitinkamai *SentiII* emocinio intensyvumo viršutinė riba.

Tarkime, turimas įrašą, kurio $score_{Google} = -0.8$ bei $magnitude_{Google} = 2$ (t.y. emocinis įvertis labai neigiamas), tai jo $SO_{SentiII} = -2.4$, o tai irgi žymi labai neigiamą SO įvertį. Gali būti tokių atvejų, kai $magnitude_{Google} > 2$ ir paskaičiuotas $SO_{SentiII} < \pm 3$. Tokiu atveju emocinį įvertį normalizuosime į $SO_{SentiII} = \pm 3$. Vadovaujantis šia logika buvo perrašyta 5

lentelė, papildžius *SentiII* SO įverčiO stulpeliu. Taigi 6 lentelė reprezentuoja emotikonų modelių papildymą, kuris buvo pritaikytas *SentiII* algoritmui. Čia skaičiavimuose buvo teigta, kad $\textit{magnitude}_{Google} = \textit{magnitude}_{virs} = 2$.

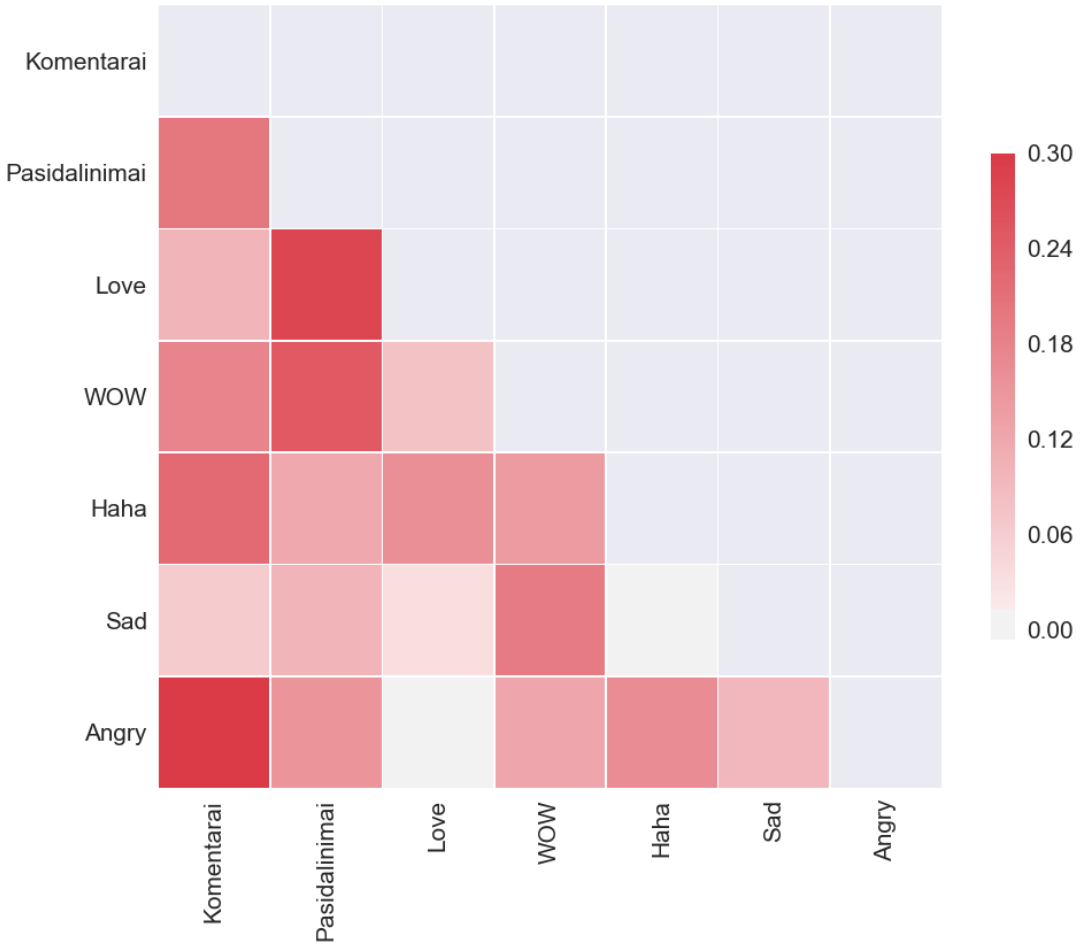
6 lentelė. Emotikonų kombinacijų suteikiama sentimentinė informaciją ir jų modeliai.

Emotikonų Kombinacija	$\textit{score}_{Google}$	$SO_{SentiII}$*SentiII*
„Angry“ – „Haha“	$[-1\ ,\ -0.5]$	$[-3\ ,\ -1.5]$
„Angry“ – „Love“	–	–
„Angry“ – „WOW“	$[-0.5\ ,\ 0.0]$	$[-1.5\ ,\ 0.0]$
„Angry“ – „Sad“	$[-1\ ,\ -0.5]$	$[-3\ ,\ -1.5]$
„Haha“ – „Love“	$[0.0\ ,\ 0.4]$	$[0.0\ ,\ 1.2]$
„Haha“ – „WOW“	$[-0.7\ ,\ -0.3]$	$[-2.1\ ,\ -0.9]$
„Haha“ – „Sad“	–	–
„Love“ – „WOW“	$[0.4\ ,\ 1]$	$[1.2\ ,\ 3]$
„Love“ – „Sad“	$[0.1\ ,\ 0.5]$	$[0.3\ ,\ 1.5]$
„WOW“ – „Sad“	$[-0.5\ ,\ -0.3]$	$[-1.5\ ,\ -0.9]$

4.4. Kiti „Facebook“ elementai

Šiame skyriuje bus trumpai paanalizuoti kiti „Facebook“ platformoje esami elementai: komentarų (angl. *comment*) skaičius bei įrašo pasidalinimų (angl. *share*) kiekis, galintys suteikti sentimentinės informacijos. Tam buvo nubraižyta Pirsono (angl. *Pearson*) koreliacijos matrica, kur abejose ašyse atidėti komentarų bei pasidalinimo elementai kartu su emotikonais, o spalvų intensyvumu nurodytas Pirsono koeficientas (žr. 10 iliustraciją). Pastebima, kad įrašai su daugiausia komentarų sulaukia „Angry“ bei „Haha“ reakcijų. Tokie tekstai dažniausiai yra kontraversiški ir sukuria diskusijas bei neapykantą. Tuo tarpu įrašai, kuriais pasidalina daugiausiai, turi „Love“ bei „WOW“ reakcijas, kadangi žmonės mieliau parodys kitiems juokingus ar neįtikėtinus dalykus nei negatyvius. Taip pat verta paminėti, kad „Love“ – „Angry“ bei „Haha“ – „Sad“ emotikonų deriniai turi beveik nulinę koreliaciją, o tai jau buvo pastebėta ir aptarta 4.3 poskyryje.

Taigi, komentarų bei pasidalijimo kiekis gali padėti įvardyti teksto poliariškumą, tačiau yra sunku nustatyti, kada šių elementų kiekis yra pakankamai didelis, kad būtų galima padaryti tokią išvadą.



10 pav. Koreliacijos matrica su įvairiais „Facebook“ elementais.

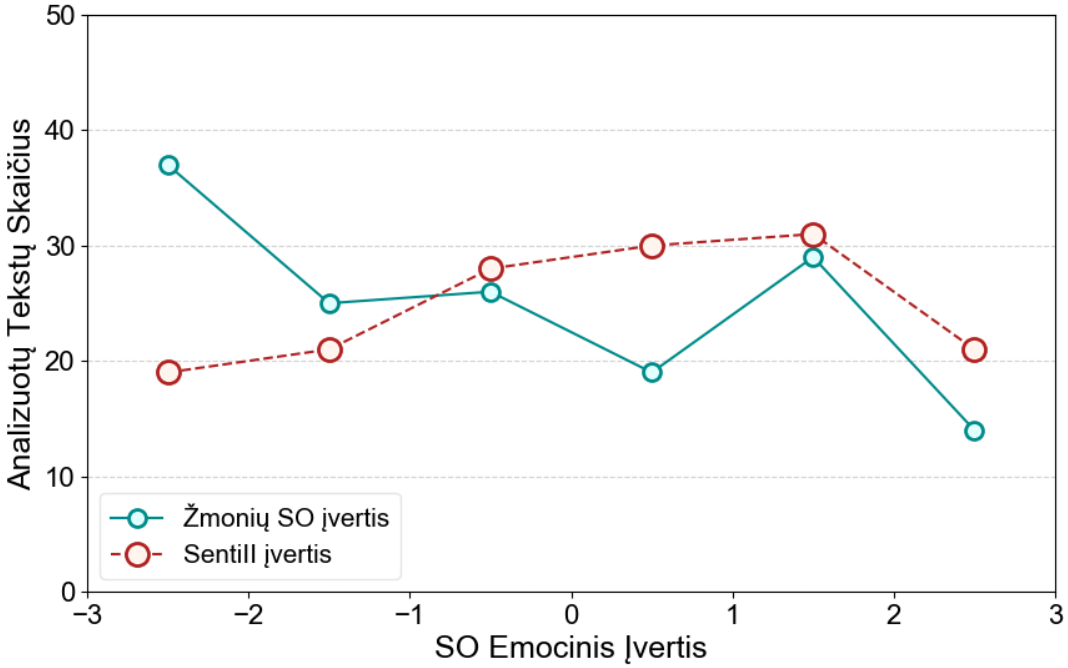
Dėl to šiame darbe nenaudojami minėti aspektai kaip galimas sentimentinės analizės informacijos šaltinis.

5. *SentiII* papildymas emotikonų modeliais

Didžiausi sentimentinio uždavinio sprendimo iššūkiai yra meninių raiškos priemonių neidentifikavimas (pavyzdžiui, oksimoronai, ironija, palyginimai), nesugebėjimas įvertinti sakinio konteksto esant šių dienų aktualijoms bei praeities konteksto neturėjimas, kuris gali būti aktualus analizuojamam įrašui. Šiame skyriuje bus aptarta, kaip naudojantis pagalbiniu parametru – emotikonų teikiama sentimentine informacija – galima pagerinti *SentiII* tikslumą.

5.1. *SentiII* be emotikonų modelių

Kaip jau minėta 3.2 skyriuje, 5 nepriklausomi ekspertai peržiūrėjo 150 įrašų, pasirinktų iš lietuviškų „Facebook“ puslapių, bei pateikė teksto SO įvertį skalėje nuo -3 (labai neigiamas) iki 3 (labai teigiamas). SO įverčio balas nebūtinai turėjo būti sveikas skaičius. Tie patys tekstai išanalizuoti su *SentiII* algoritmu naudojant 3.2 skyriuje nustatytus optimalius parametrus, tai yra kai sutapimo toleruojamas koeficientas lygus $\eta = 70\%$ bei sakinių susietumo parametras γ yra pusinis normalusis skirstinys. Rezultatai buvo palyginti tarpusavyje ir pateikti 11 paveikslėlyje. Galima



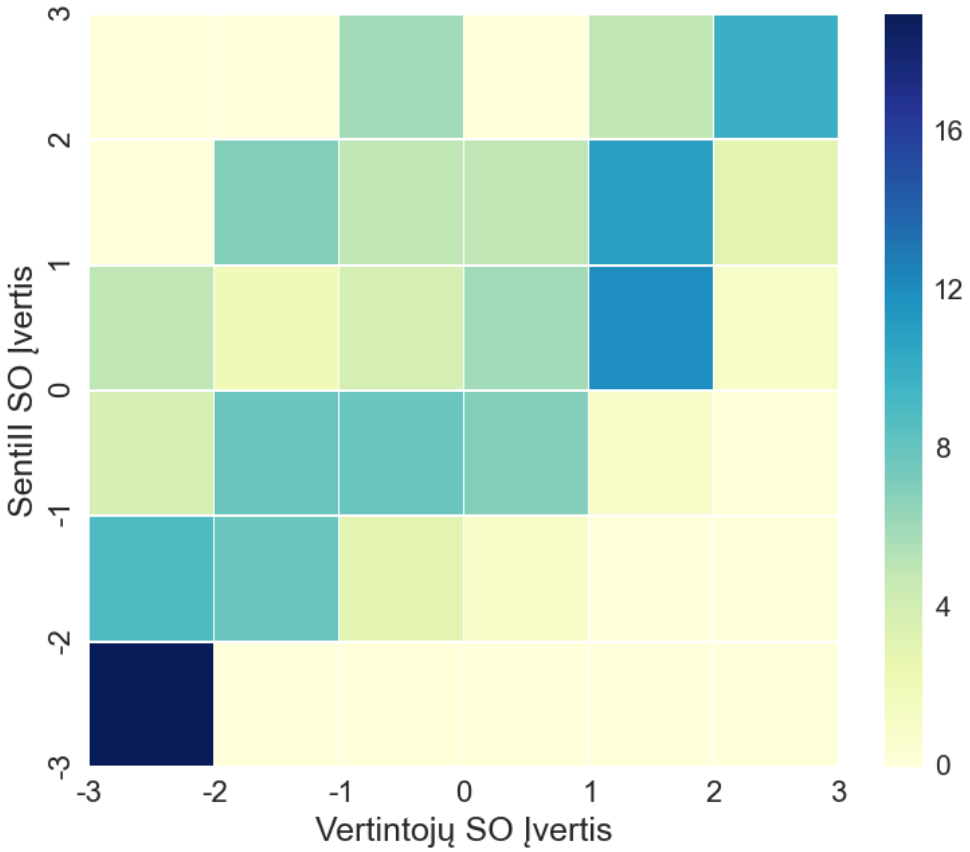
11 pav. Tekstų kiekio pasiskirstymo diagrama pagal *SentiII* (brūkšniuota linija) ir vertintojų (vientisa linija) suteiktą SO įvertį.

pastebėti, kad algoritmui buvo sunku susidoroti su kraštutiniais SO įverčiais, o tai gali būti paaiškinta žmogiškuoju faktoriumi – tekste minimi aspektai kiekvienam asmeniškai yra skirtingi ir jau nuo pirmų sakinių gali nulemti teksto emocinį įvertį. Pavyzdžiui, etninės grupės, kurias kiekvienas individas vertina skirtingai, gali sukelti kraštutines emocijas (t.y. labai neigiamą arba teigiamą). Taip pat pastebima, kad vertintojai didžiąją tekstų dalį sudėjo į neigiamo poliariškumo pusę, o tai yra visiškai priešinga algoritmo atveju. Didžiausias neatitikimas atsirado $SO \in [-3\ ,\ -1]$ rėžyje, kur, tikėtina, rastume ironiją ar sarkazmą. Šią dalį algoritmas galimai sudėliojo $SO \in [-1\ ,\ 2]$ rėžyje, neatpažindamas šios meninės raiškos priemonės ir registruodamas jos žodžius su teigiamu SO įverčiu. Visgi algoritmas neblogai susidorojo su $SO \in [1\ ,\ 3]$ rėžiu.

Norint geriau suprasti $SO \in [-3\ ,\ -1]$ nesutapimą bei sužinoti, į kuriuos SO įverčio rėžius algoritmas sudeda vertintojų įvertintus tekstus, buvo nubraižyta 2D histograma (žr. 12 iliustraciją). Abscisių bei ordinačių ašyse atitinkamai atidėti vertintojų SO_{vert} bei algoritmo SO įverčių $SO_{SentiII}$ rėžiai, o spalvų intensyvumu nurodytas įrašų skaičius, pakliūnantis į konkretų emocinį rėžį.

12 grafiką reikia suprasti taip, kad jei algoritmo ir vertintojų nuomonės pilnai sutaptų, tai tik grafiko įstrižainė turėtų tam tikrus atspalvius. Tuo tarpu aplinkui esantys plotai būtų baltos spalvos (žr. A priede esančią 20 iliustraciją). Iš pirmo žvilgsnio galima teigti, kad algoritmas neblogai susidorojo su $SO_{vert} \in [-3\ ,\ 0]$ rėžiu, tačiau visi šie įrašai pasiskirstė $SO_{SentiII} \in [-1\ ,\ 3]$ rėžyje. Pavyzdžiui, jei vertintojas tekstui skyrė -1.8, tai algoritmas 36% tokių tekstų skirtų teigiamą poliariškumą, o 27% atvejų sudėtų į $SO_{SentiII} \in [1\ ,\ 2]$ rėžį. Pagrindinė šio reiškinio priežastis – algoritmo nesugebėjimas identifikuoti meninių raiškos priemonių bei dienos aktualijų. Dėl šios priežasties tekstas yra suprantamas „tiesiogiai“, taip priskiriant teigiamą poliariškumą.

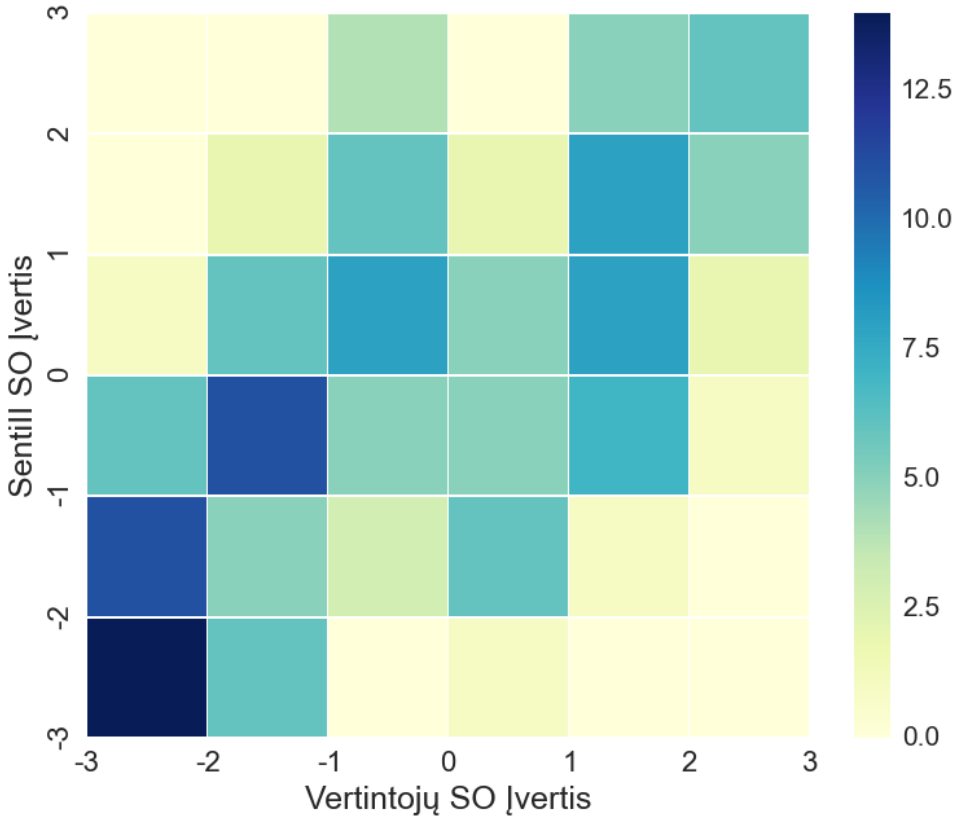
Tą patį galima pastebėti $SO_{vert} \in [1\ ,\ 2]$ rėžyje, kur dėl tekste naudojamų palyginimų algoritmas didžiąją dalį tekstų priskyrė prie silpnai teigiamo / neutralaus rėžio. Nepaisant to didžioji



12 pav. Vertintojų bei *SentiII* algoritmo SO įverčių pasiskirstymas 2D histogramoje, kur langelis reprezentuoja įrašų skaičių pagal pateiktą skalę dešinėje.

dalis įrašų iš $SO_{vert} \in [-3\ ,\ -2]$ bei $SO_{vert} \in [2\ ,\ 3]$ rėžių atitiko algoritmo įverčius. Pasigilinus pastebėta, kad karštinės poliariškumo vertės buvo lengvai identifikuojamos naudojant tik sentimentinius žodžius, kadangi tekstų autoriai retai kada naudojo meninės raiškos priemones šiems įrašams. Visų šių trūkumų sprendimai bus aptarti 5.2 skyriuje.

Dėl įdomumo galima pasinaudoti ta pačia 12 iliustracijos vizualizavimo koncepcija ir nubraižyti, kaip algoritmo SO įverčiai pasiskirstytų, jei būtų naudotas kitas sakinių susietumo parametras $\gamma = \frac{1}{d}$ (žr. 13 iliustraciją). Iš karto pastebima, kad nemaža dalis įrašų pasislinko iš grafiko įstrižainės ir tolygiai pasiskirstė per kelis rėžius, tokiu būdu sumažindami atitikimą tarp vertintojų bei algoritmo. Pavyzdžiui, jei vertintojas tekstui skirtų 1.8, tai algoritmas $\simeq 25\%$ tokių atveju sudėtų į $[0\ ,\ 1]$ arba į $[1\ ,\ 2]$, arba į $[2\ ,\ 3]$ rėžius – visa tai yra labai neapibrėžta. Tokį rezultatą galima paaiškinti tuo, kad toks sakinių susietumas neturi emocinės atminties (žr. 3.2 skyrių). Jeigu teksto pradžioje SO įvertis buvo labai neigiamas, o gale silpnai teigiamas, galutinis SO įvertis taps neutralus arba net silpnai teigiamas, priklausomai nuo teksto dydžio. Tai lemia 13 grafike esantį SO įverčių pasiskirstymą per visą plotą, išskyrus kraštines vertes (t.y. -3 bei $+3$), kurių įrašai dažniausiai nekeičia emocinio įverčio per visą tekstą.

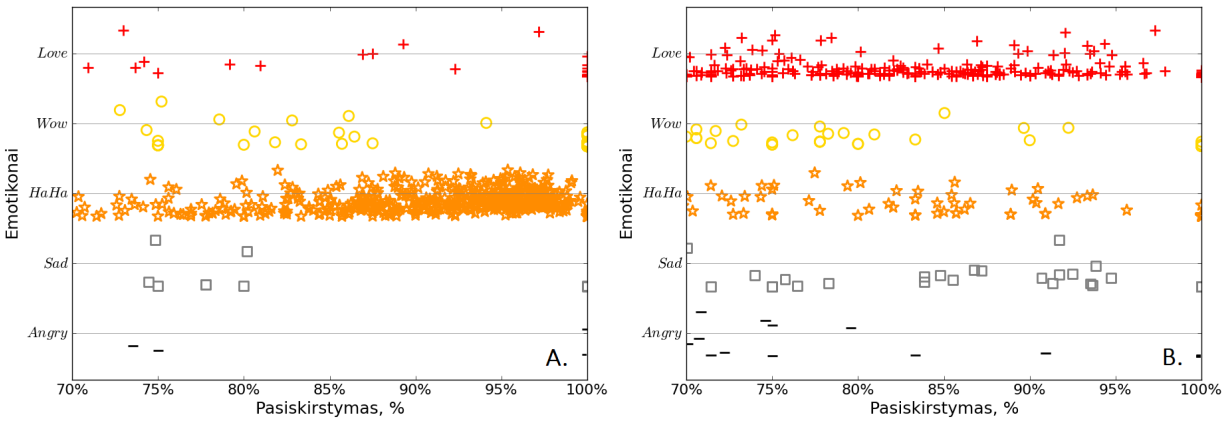


13 pav. Vertintojų bei *SentiII* algoritmo SO įverčių pasiskirstymas 2D histogramoje, kur langelis reprezentuoja įrašų skaičių pagal pateiktą skalę dešinėje, kai $\gamma = \frac{1}{d}$.

5.2. *SentiII* su emotikonų modeliais

Prieš pritaikant 6 lentelėje apibrėžtus emotikonų modelius, įsitikinta, ar lietuviškuose „Facebook“ puslapiuose emotikonų procentinis pasiskirstymas gali nusakyti to puslapio turinį. Pasinaudojant 4 iliustracijos grafine koncepcija, 14 iliustracijoje atvaizduotas „Geriausios Demotyvacijos“ (satyros bei humoro turinio puslapis) bei „Delfi.lt“ (naujienų bei šių dienų aktualijų turinio puslapis) grupių reakcijų procentinis pasiskirstymas, naudojant tik 2017 metų įrašus. Matoma, kad didžiausią „Geriausios Demotyvacijos“ (14 paveiksliukas, A dalis) grupės emotikonų profilio dalį sudaro „Haha“ reakcija, kai kitos, ypač „Angry“ ir „Sad“, turi vos kelis įrašus. Tai visiškai atspindi šios grupės turinį. Tuo tarpu naujienų portalas „Delfi.lt“ (14 paveiksliukas, B dalis) turi visą emotikonų spektrą, o tai būdinga, kai grupėje aptariamos kontraversiškos aktualijos bei įvykiai. Taigi abiejų grupių reakcijų profiliai puikiai atspindi jų turinius, o palyginus su 4 iliustracijoje pavaizduotomis angliškų grupių atitikmenimis galima teigti, kad emotikonų turima sentimentinė informacija išlieka nepaisant demografinės aplinkos.

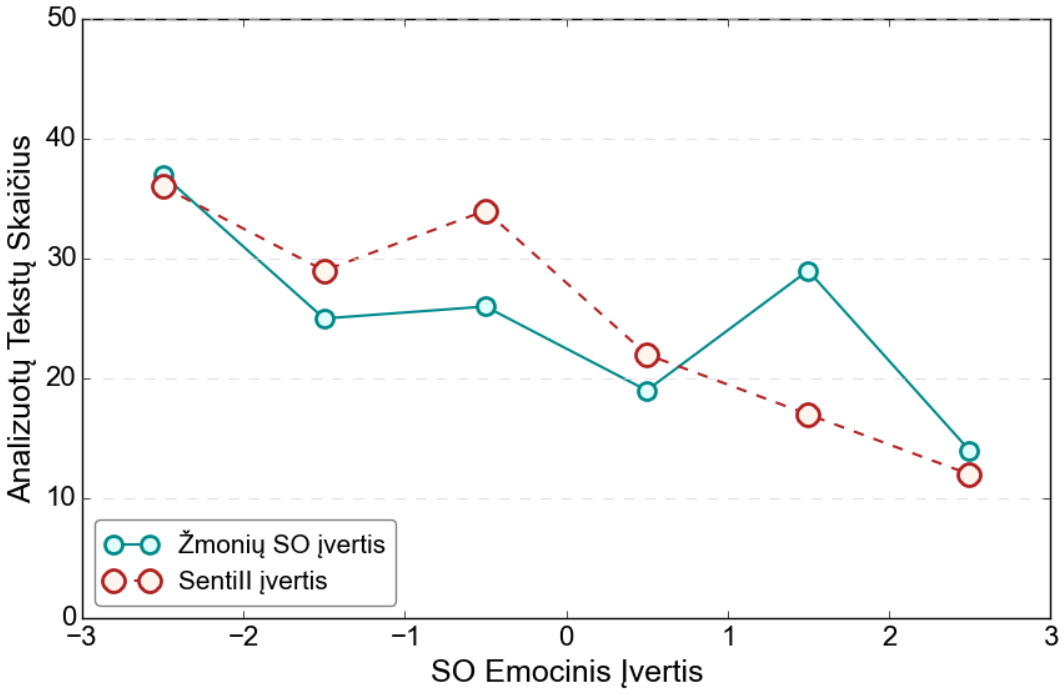
Įsitikinus, kad 6 lentelėje apibendrinti modeliai, pagrįsti emotikonų procentiniu pasiskirstymu, gali veikti taip pat veiksmingai su lietuviškais tekstais kaip ir su angliškais, bus aprašytas algoritmo papildymas 4 pseudokodu. Taigi *SentiII* algoritmu gavus SO_{tekst} įvertį, tikrinta ar šis įvertis patenka į atitinkamo emotikonų modelio rėžį $SO_{tekst} \in emo\_vert_{sak}$. Jei taip, veiksmai baigiami. Priešingu atveju vertinama, ar algoritmo SO įvertis yra to pačio poliariškumo kaip numatomo emotikono modelio. Tokiu būdu įvertinama, kaip stipriai algoritmas apsiriko palyginus su emotiko-



14 pav. „Facebook“ emotikonų kiekio pasiskirstymas per įrašą dvejose skirtingose „Facebook“ grupėse, kur A. vaizduoja „Geriausios Demotyvacijos“ grupės įrašus ir B. „Delfi.lt“ grupės įrašus.

no modelio nešamu įverčiu. Jei emocinis poliariškumas sutampa, tuomet galutinis SO_{tekst} įvertis bus lygus reakcijų modelio rėžio viduriui. Tokiu atveju, jei poliariškumas nesutampa (pavyzdžiui, algoritmas identifikuoja tekstą, turintį teigiamą SO įvertį, o atitinkamas emotikonų modelis neigiamą), skaičiuojamas algoritmo įverčio emocinis stiprumas $\frac{SO_{SentiII}}{3}$, kuris naudojamas kaip daugiklis reakcijų modelio nurodytame rėžyje emo_vert_{sak} vietos įvertinimui.

Įdiegus emotikonų modelių papildymą į *SentiII* algoritmą, dar kartą surinkti 150 tekstų SO įverčiai ir palyginti su turimais vertintojų įverčiais. Rezultatai pateikti 15 iliustracijoje naudojant



15 pav. Tekstų kiekio pasiskirstymo diagrama pagal *SentiII* su emotikonų modeliais (brūkšniuota linija) ir vertintojų (vientisa linija) suteiktą SO įvertį.

tą patį vizualizavimo formatą kaip ir 11 iliustracijoje. Galima pastebėti, kad algoritmas neblogai susitvarkė su SO įverčio kraštinėmis vertėmis, o jo gauta bendra kreivė per visą $[-3\ ,\ 3]$ rėžį

4 algoritmas. Emotikonų modelių papildimo įgyvendinimo pseudokodas.

Įvestis: SO_{tekst}: teksto SO įvertis, pradžioje tai *SentiII* gautas įvertis; emo_{tekst}: didžiausią įrašo visų emotikonų dalį sudaranti pora; $emo\_mod = [\prime angry - haha\prime, \prime haha - love\prime, ...]$: emotikonų modelių pavadinimų masyvas; $emo\_vert = [\prime -1.5\ ,\ -3\prime, \prime 0\ ,\ 1.2\prime ...]$: atitinkamai emotikonų modelių SO įverčių rėžiai, kur kraštai $emo\_vert^0 = -1.5$, $emo\_vert^1 = -3$; emo_polar: emotikonų modelių poliariškumas

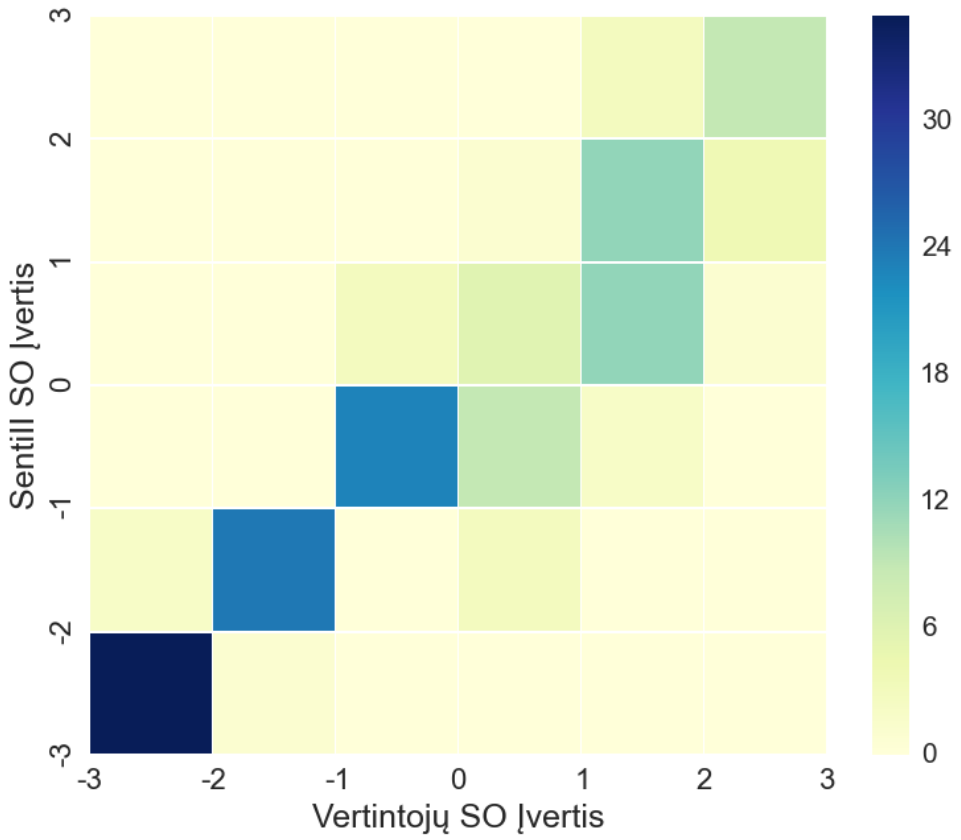
Išvestis: SO_{tekst}: galutinis teksto SO įvertis

$emo\_vert_{sak} \leftarrow emo\_vert[emo\_mod.index(emo_{tekst})]$ pasirenkamas tinkamas modelis ir jo rėžiai pagal sakinį
if SO_{tekst} **in** emo_vert_{sak} **then**
    PASS
else
    tikriname ar *SentiII* SO įvertis tokio pačio poliariškumo kaip ir pritaikytas emotikonų modelis
    if $SO_{tekst}^{poliariskumas} = emo\_polar$ **then**
        $SO_{tekst} \leftarrow emo\_vert_{sak}^{min} + 0.5 \times (emo\_vert_{sak}^{1} - emo\_vert_{sak}^{0})$
    else
        $SO_{tekst} \leftarrow emo\_vert_{sak}^{min} - (SO_{tekst}/3) \times |emo\_vert_{sak}^{1} - emo\_vert_{sak}^{0}|$
    end if
end if
return SO_{tekst}

yra labai panaši į vertintojų įverčių pasiskirstymą. Šį kartą matoma, kad turimas teigiamų įverčių trūkumas, kurio dalis nukeliavo į silpnai neigiamus / neutralius rėžius.

Buvo nubraižyta 2D histograma (žr. 16 iliustraciją), kur kaip ir anksčiau abscisių bei ordinačių ašyse atidėti vertintojų SO_{vert} bei algoritmo SO įverčių $SO_{SentiII}$ rėžiai, o spalvų intensyvumu nurodytas įrašų skaičius, pakliūnantis į konkretų emocinį rėžį. Matoma, kad *SentiII* su labai maža paklaida įvertino visą $SO_{vert} \in [-3\ ,\ -1]$ rėžį panašiai kaip vertintojai. Tai rodo, kad emotikonų modeliais pavyko išspręsi meninių raiškos priemonių problemą. Visgi matoma, kad ties $SO_{vert} \in [0\ ,\ 3]$ rėžiu, atitikimo paklaida yra kiek didesnė, nors palyginus su $SO_{SentiII}$ versija be reakcijų modelių (žr. 12), atrodo kur kas geriau. Tai galima paaiškinti prisiminus emotikonų modelių veikimo rėžius, atvaizduotus 9 iliustracijoje. Kaip matoma neigiamo poliariškumo rėžis turi net 5 modelius, kurie persikloja iki trijų kartų, tuo tarpu teigiamame ruože turime tik 3 modelius, iš kurių tik vienas apibūdina $[1.5\ ,\ 3]$ rėžį. Tokiu būdu, jei algoritmo išvestas SO įvertis nepatenka į emotikonų modelio, pavyzdžiui, „Love“ – „WOW“ būdingą rėžį, naujas SO įvertis turės gan didelį rėžį būti įdėtas pagal 4 algoritmą. Šis neapibrėžtumas lėmė 15 iliustracijoje matomą platų $SO_{SentiII}$ įverčių pasiskirstymą esant konkrečiam teigiamam SO_{vert} įverčiui.

Galima drąsiai teigti, kad „Facebook“ emotikonų modeliai padėjo sumažinti neatitikimą tarp *SentiII* algoritmo bei vertintojų duotų SO įverčių. Vos $\simeq 25\%$ visų analizuotų tekstų SO įverčiai buvo skirtingi, laikantis vieno balo SO įverčio rėžiu. Tuo tarpu algoritmo versija be emotikonų papildinio turėjo net $\simeq 51\%$ įrašų neatitikimą su vertintojų nuomone.



16 pav. Vertintojų bei *SentiII* su emotikonų modeliais algoritmo SO įverčių pasiskirstymas 2D histogramoje, kur langelis reprezentuoja įrašų skaičių pagal pateiktą skalę dešinėje.

Tuo tarpu algoritmo versija be emotikonų papildinio turėjo net $\simeq 51\%$ įrašų neatitikimą su vertintojų nuomone.

6. Išvados

Šiame tiriamajame darbe buvo sėkmingai sukurtas sentimentų žodynu pagrįstas SO emocinio lygio įvertinimo algoritmas *SentiII*. Buvo nustatyta, kad pusinis normalusis skirstinys gali geriausiai apibūdinti sakinių SO įverčių susietumą, o geriausias algoritmo tikslumas $F_1 = 0.751$ pasiekiamas naudojant 70% kaip toleruojamą dviejų žodžių frazių sutapimo koeficientą Levenšteino atstumui skaičiuoti.

Tiriant daugiau nei kelis milijonus vartotojų turinčias „Facebook“ grupes, įsitikinta, kad emotikonai suteikia sentimentinės informacijos apie įrašą. Remiantis šiuo rezultatu buvo sukurti 10 emotikonų modelių, galinčių tiksliai įvertinti teksto poliariškumą bei semantinio įverčio rėžį. Vėliau *SentiII* algoritmas buvo papildytas šiais „Facebook“ reakcijų pagrįstais modeliais.

Algoritmo testavimas buvo atliekamas pateikiant socialiniame tinkle „Facebook“ publikuojamus tekstus iš kelių populiariausių lietuviškų puslapių. Išrinkti 150 tekstų buvo įvertinti 5 nepriklausomų ekspertų ir palyginti su algoritmo (su ir be emotikonų papildymu) suteiktais SO įverčiais. Nustatyta, kad *SentiII* su emotikonų modeliais pasirodė 2 kartus geriau nei pagrindinė jo versija, įvertinant sarkazmą, menines raiškos priemones bei dienos aktualijas.

7. Darbo gairės

Siekiant pagerinti *SentiII* algoritmo veikimą bei pritaikymą, bus peržvelgti keli svarbiausi galimi tobulinimo aspektai. Visų pirma, emotikonų porų modeliai neidealiai įvertina teigiamą SO įverčio rėžį, kadangi nemaža šio rėžio dalis perdengiama tik vienu modeliu. Trijų reakcijų porų modeliai galėtų suskaidyti rėžį į mažesnes dalis, taip sukonkretindami įrašo SO įvertį. Šis sukonkretinimas galėtų veikti tokiu būdu – išrenkamos trys didžiausią įrašo dalį sudarančios reakcijos, kur pirmos dvi nurodytų poliariškumą bei SO rėžį, o trečia konkretintų ties kuria rėžio kraštine verte galėtų būti teksto įvertis.

Darbe turėtų būti atlikta platesnė pasirinktos vertintojų grupės analizė. Skiriant daugiau dėmesio žmonių imties sudarymui bei atsižvelgiant į jų emocinį intelektą, būtų sukuriamas jų emocinis modelis. Tokiu būdu, testuojant algoritmo veikimą, būtų galima nuspėti ekstremalius SO įverčių nuokrypius tam tikra tematika. O tai leistų geriau suprasti rezultatus bei vėliau *SentiII* palyginti su kitais egzistuojančiais emocijų vertinimo modeliais.

Tolimesniais tyrimais butų galima konkretinti įrašų perteikiamas emocijas, pavyzdžiui identifikuojant tekste laimę, baimę ar pyktį, ir įvertinant jų stiprumo lygį. Būtų įdomu panagrinėti, ar yra tam tikras šių emocijų sąryšis su „Facebook“ emotikonų reakcijomis. Tokios sąsajos galėtų tikslinti algoritmo rezultatus. Visus įvardintus patobulinus galima būtų pritaikyti analizuojant Lietuvos naujienų portalus ar tas pačias „Facebook“ grupes, suprantant publikuojamą turinį, jo kaitą laike bei skaitytojus.

Literatūros šaltiniai

[1] Laikykitės ten, 2017.
https://www.facebook.com/pg/laikykitesten.

[2] Facebook graph api, 2018.
https://developers.facebook.com/docs/graph-api/.

[3] Google natural language api, 2018.
https://cloud.google.com/natural-language/docs/analyzing-sentiment.

[4] Rel Guzman Apaza, José Eduardo Ochoa Luna, Laura Vanessa Cruz Quispe, and Elizabeth Vera Cervantes. Predicting reactions to blog headlines. pages 43–47, 2016.

[5] Anthony Aue and Michael Gamon. Customizing sentiment classifiers to new domains: A case study. 1(1-3):2–1, 2005.

[6] Angelo Basile, Tommaso Caselli, and Malvina Nissim. Predicting controversial news using facebook reactions. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017), Rome, Italy, December 11-13, 2017.*, 2017.

[7] Ryan Bozeman. Here's what happens every second on the internet, 2016.
https://www.scribblrs.com/heres-happens-every-second-internet/.

[8] BrightLocal. Local consumer review survey, 2016.
https://www.brightlocal.com/learn/local-consumer-review-survey/.

[9] Lindsey Brylow. Emotional aspects of facebook textual posts a framework for marketing researches. *European Journal of Science and Theology*, 12(6):187–197, 2016.

[10] Lindsey Brylow. Facebook reactions, 2016.
https://www.ogilvypr.com/wp-content/uploads/2016/03/Facebook_Reactions-OPR-2.26.16.pdf.

[11] Yejin Choi and Claire Cardie. Learning with compositional semantics as structural inference for subsentential sentiment analysis. pages 793–801, 2008.

[12] Thomas Dimson. Emojineering: Machine learning for emoji trends by instagram, 2015.
http://instagram-engineering.tumblr.com/post/117889701472/emojineering-part-1-machine-learning-for-emoji.

[13] Emilio Ferrara, Onur Varol, Filippo Menczer, and Alessandro Flammini. Traveling trends: social butterflies or frequent fliers? In *Conference on Online Social Networks, COSN'13, Boston, MA, USA, October 7-8, 2013*, pages 213–222, 2013.

[14] Emilio Ferrara and Zeyao Yang. Quantifying the effect of sentiment on information diffusion in social media. *PeerJ Computer Science*, 1:e26, 2015.

[15] Qi Ge, Alexader Kurov, and Marketa Halova Wolfe. Stock market reactions to presidential social media usage: Evidence from company-specific tweets, 2017.
http://lamacro.davidson.edu/wp-content/uploads/gravity_forms/

1-95b74353d944e7d67926fb1900938dee/2017/05/Ge-Kurov-Wolfe_2017_
Stock-Market-Reactions-to-Presidential-Social-Media-Usage.pdf.

[16] Steven L. Heston and Nitish R. Sinha. News versus sentiment: Predicting stock returns from news stories. 2015.
http://dx.doi.org/10.17016/FEDS.2016.048.

[17] Sara Hofmann, Daniel Beverungen, Michael Räckers, and Jörg Becker. What makes local governments' online communications successful? insights from a multi-method analysis of facebook. *Government Information Quarterly*, 30(4):387–396, 2013.

[18] Syed Akib Anwar Hridoy, M. Tahmid Ekram, Mohammad Samiul Islam, Faysal Ahmed, and Rashedur M. Rahman. Localized twitter opinion mining using sentiment analysis. *Decision Analytics*, 2:8, 2015.

[19] David A. Huffaker and Sandra L. Calvert. Gender, identity, and language use in teenage blogs. *Journal Computer-Mediated Communication*, 10(2), 2005.

[20] Haruna Isah, Paul R. Trundle, and Daniel Neagu. Social media analysis for product safety using text mining and sentiment analysis. pages 1–7, 2014.

[21] Ryan Kelly and Leon Watts. Characterising the inventive appropriation of emoji as relationally meaningful in mediated close personal relationships. *Experiences of Technology Appropriation: Unanticipated Users, Usage, Circumstances, and Design*, 2015.

[22] Peter Kim, Elana Anderson, and Jennifer Joseph. The forrester wave: Brand monitoring, q3 2006. *Forrester Wave (white paper)*, 2006.

[23] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue B. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010*, pages 591–600, 2010.

[24] Petra Kralj Novak, Jasmina Smailovic, Borut Sluban, and Igor Mozetic. Sentiment of emojis. *CoRR*, abs/1509.07761, 2015.

[25] Bo Pang, Lee Lillian, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. pages 79–86, 2002.

[26] Livia Polanyi and Annie Zaenen. Contextual valence shifters. pages 1–10, 2006.

[27] Hassan Saif, F. Javier Ortega, Miriam Fernández, and Iván Cantador. Sentiment analysis in social streams. In *Emotions and Personality in Personalized Services - Models, Evaluation and Applications*, pages 119–140. 2017.

[28] Franco Salvetti, Christoph Reichenbach, and Stephen Lewis. Opinion polarity identification of movie reviews. pages 303–316, 2006.

[29] Antonios Siganos, Evangelos Vagenas-Nanos, and Patrick Verwijmeren. Divergence of sentiment and stock market trading. *Journal of Banking & Finance*, 78:130–141, 2017.

[30] Maite Taboada, Caroline Anthony, and Kimberly D. Voll. Methods for creating semantic orientation dictionaries. pages 427–432, 2006.

[31] Justas Tamašauskas and Linas Bukauskas. Skaitmeninio teksto sentimentų vertinimo metodų tyrimas, 2017.

[32] Danny Thomson and Eric Ehizokhale. Analysing social network reactions to 2016 republican primaries, 2015.
http://snap.stanford.edu/class/cs224w-2015/projects_2015/Analysing_Social_Network_Reactions_to_2016_Republican_Primaries.pdf.

[33] Ye Tian, Thiago Galery, Giulio Dulcinati, Emilia Molimpakis, and Chao Sun. Facebook sentiment: Reactions and emojis. *SocialNLP 2017*, page 11, 2017.

[34] Richard M. Tong. An operational system for detecting and tracking opinions in on-line discussion. In *Working Notes of the ACM SIGIR 2001 Workshop on Operational Text Classification*, volume 1, pages 1–6, 2001.

[35] Mikalai Tsytsarau, Themis Palpanas, and Malú Castellanos. Dynamics of news events and social media reaction. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, pages 901–910, 2014.

[36] Diego Tumitan and Karin Becker. Tracking sentiment evolution on user-generated content: A case study on the brazilian political scene. In *XXVIII Simpósio Brasileiro de Banco de Dados - Short Papers, Recife, Pernambuco, Brasil, September 30 - October 3, 2013.*, pages 24:1–24:6, 2013.

[37] Peter D. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. pages 417–424, 2002.

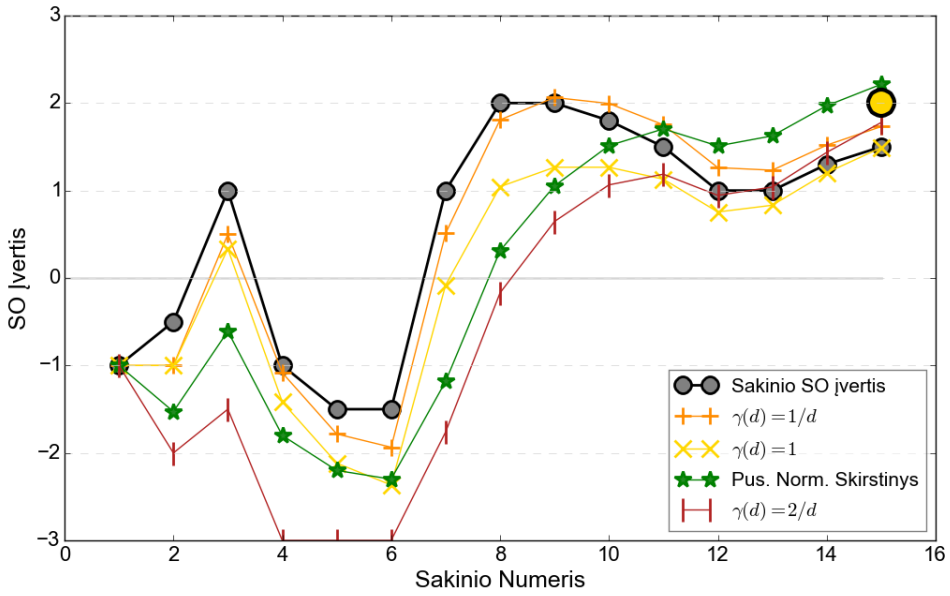
[38] Peter D. Turney and Michael L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Trans. Inf. Syst.*, 21(4):315–346, 2003.

[39] Casey Whitelaw, Navendu Garg, and Shlomo Argamon. Using appraisal groups for sentiment analysis. pages 625–631, 2005.

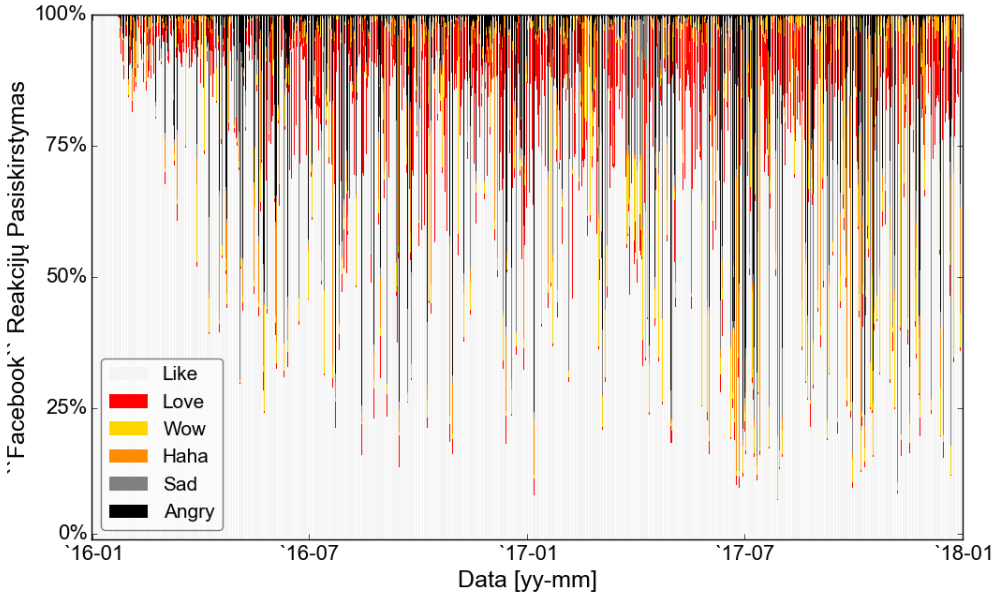
[40] Janyce Marbury Wiebe. Recognizing subjective sentences: A computational investigation of narrative text. 1990.

Priedai

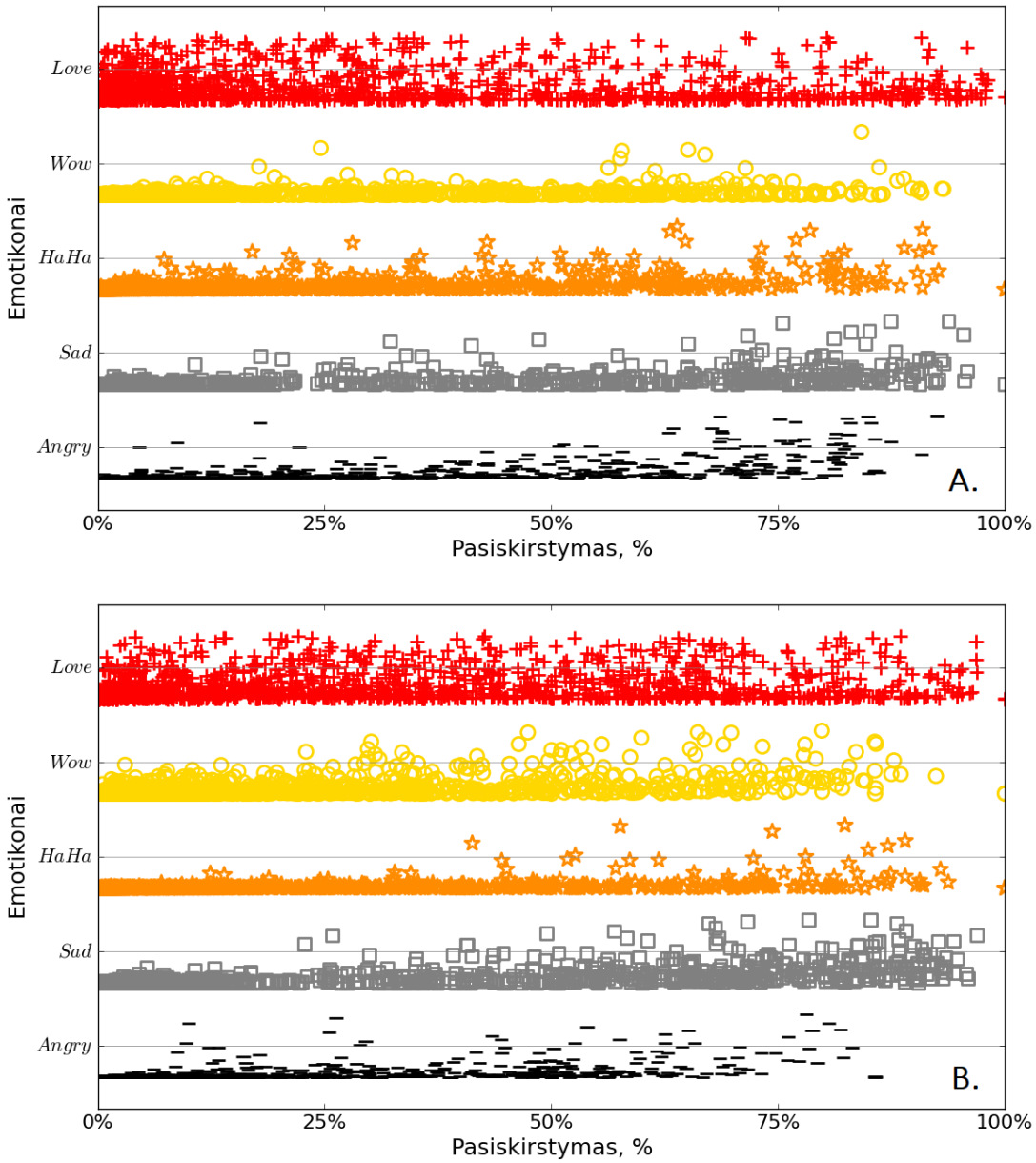
A. Papildomi grafikai



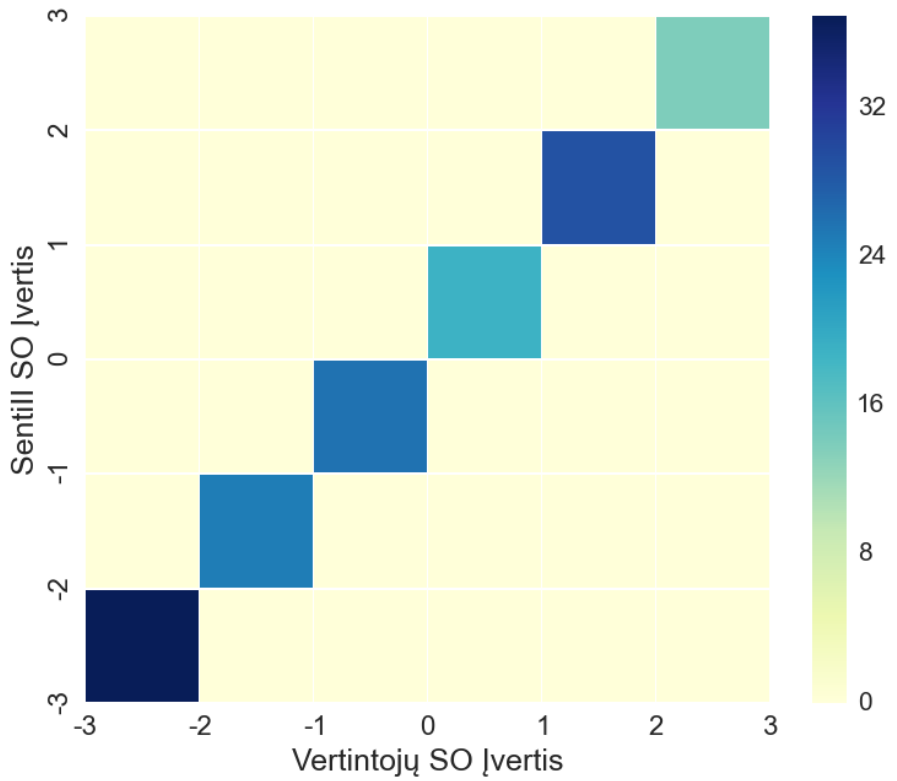
17 pav. Dar vieno teksto SO slenkančio įverčio grafikas, kuriame atvaizduojamas skirtingos $\gamma(d)$ strategijos. Iliustracijos dešinėje esantis didelis burbulas žymi vertintojų skirtą bendrą (t.y. visų sakinių) teksto SO įvertį.



18 pav. „Facebook“ reakcijos emotikonų kiekio pasiskirstymas per dienos įrašus „Facebook“ puslapyje „Foxnews“ nuo 2016 metų iki 2018 pradžios.



19 pav. „Facebook“ emotikonų kiekio pasiskirstymas per įrašą dviejose skirtingose „Facebook“ naujienų grupėse, kur A. vaizduoja „CNN“, o B. – „BBC“ įrašus.



20 pav. Vertintojų bei *SentiII* algoritmo SO įverčių pasiskirstymas 2D histogramoje, tuo atveju jei SO įverčiai idealiai sutaptų.

B. *SentiII* naudojimosi instrukcijos be ir su „Facebook“ emotikonų papildimu

PAKETO STRUKTŪRA

- **dic_words.dat** – failas su sentimentų žodynu, bei SO įverčiais (nuo -3 iki 3);
- **dic_int.dat** -– failas su SO silpninančiais / stiprinančiais žodeliais bei jų stiprinimo, bei silpninimo koeficientais;
- **test0.dat**, **test1.dat**, **test2.dat** -– failai su pavyzdiniu tekstu algoritmui apdoroti;
- **Senti_main.py** — pagrindinis programos kodas. Kode naudojamos bibliotekos yra jau įrašytos standartiniame Python pakete;
- **fb_read.py** — programos kodas renkantis tekstus bei susijusią informaciją iš pasirinktų „Facebook“ grupių. Tikėtina, kad vartotojui reikės įsidiegti kode naudojamą biblioteką *urllib2*, nors ši biblioteka pagal numatymą yra standartiniame naujose Python versijose. Šis kodas yra iššaukiamas **Senti_main.py** kodo pagal nustatymus.

PALEIDIMAS

Kodas paleidžiamas atsidarius Senti_main.py per bet kokį Python kalbos redaktorių ir spaudžiant „RUN“ arba paleidžiant per terminalą su komanda *python Senti_main.py*.

KODO KONFIGŪRACIJA

Algoritmo nustatymai bei analizuojamų tekstų pasirinkimai keičiami Senti_main.py kode.

- 243–244 eilutėse nurodoma žodynų vieta;
- 247–251 eilutėse nurodomas inversijos poslinkio dydis, sakinių susietumo logika, toleruojamas dviejų frazių sutapimo koeficientas, po kiek neutralių žodžių SO įvertis bus skaičiuojamas naujai bei ar naudoti emotikonų modelius;
- 254–262 eilutėse nurodoma „Facebook“ grupė algoritmui ištirti. Pateikėme pora lietuviškų puslapių pavyzdžių;
- 264–265 eilutėse nurodomas laiko periodas, kuriuo remiantis, algoritmas išfiltruos tik tinkamus įrašus;
- 268 eilutėje nurodomas testinis atvejis, norint išbandyti algoritmą. Jei ši eilutė nėra užkomentuota, tuomet eilučių 252–262 nustatymai bus ignoruojami.

VARTOTOJO GRAFINĖ SĄSAJA

Kodas per terminalą suteikia minimalią sąsaja su vartotoju. Ši informacija bus pateikiama vartotoju terminale:

- Nurodomas Analizuojamas tekstų šaltinis
  >Analizuojamas Šaltinis: test1.dat

- Pateikiamas analizuojamas tekstas
  !Nr.1 TEKSTAS: berečių pažeistos smegenys... tvartelio melas viską jau padarė. pasaulis žino kas pradėjo žudynes irake , sirijoje , libijoje...

- Kiekvienas teksto sakinys pateikiamas atskirai tokia tvarka: pirmoje eilutėje sakinio normalizuota forma, antroje transformuota forma identifikavus sentimentinius elementus, trečioje sakinio SO skaičiavimo formulė su SO įverčiu.
  > SAKINYS: pasaulis zino kas pradejo zudynes irake sirijoje libijoje dgstsk
  Apdorotas Sakinys: pasaulis zino kas pradejo $-2\backslash z/$ $-1\backslash z/$ $-1\backslash z/$ $-1\backslash z/$ $0.3\backslash i/$
  Skaičiavimo Formulė: -2++-1++-1++-1+(1+0.3) Rez: -1.325*

- Pateikiamas *SentiII* paskaičiuotas SO įvertis
  !Nr.1 Teksto SO Įvertis: -3

- Jei buvo naudoti emotikonų modeliai, pateikiamas naudotas modelis bei jo pakoreguotas SO įvertis. Šiuo atveju *SentiII* įvertinta SO vertė patenka į „Angry“ – „Sad“ SO rėžį.
  !Nr.1 Teksto SO Įvertis (Su "Angry-Sad" EMO Modeliu): -3

- Atskiras kiekvienam sakiniui suskaičiuotas SO įvertis
  Sakinių SO masyvas: [-1.3, -1.25, -1.325]

- Einamasis SO įvertis naudojant pasirinkta sakinių susietumo priklausomybę
  Slenkantis SO masyvas: [-1.3, -2.5860382036112037, -3]

C. Emotikonų modelio pavyzdžiai

7 lentelė. Pavyzdžiai rodantys kaip veikia emotikonų modelis

Analizuojamas sakinys	$SO_{SentiII}$ įvertis	Emotikonų modelis	Galutinis SO_{sak}
Maistas buvo *sugedęs*$^{-2}$ ir *netinkamas*$^{-1}$ vartoti.	-1.5	„Angry“ – „WOW“, $[-1.5\ ,\ 0]$	-1.5
O taip, juk jis yra pats *geriausias*$^{+3}$ ir *nuostabiausias*$^{+3}$.	$+3$	„Angry“ – „Haha“, $[-3\ ,\ -1.5]$	-3
Nors oras buvo *gan*$^{*0.3}$ *prastas*$^{-1.5}$, tačiau pasirodymas buvo *nuostabus*$^{+2.5}$.	$+0.55$	„Love“ – „WOW“, $[1.2\ ,\ 3]$	$+2.1$

D. Žodynų fragmentai

8 lentelė. Abiejų žodynų fragmentai

Sentimentų žodynas	Silpninančių / Stiprinančių žodelių žodynas
problema, -1	truputi, -0.5
nesamone, -2	siek tiek, -0.3
veblenti, -1	nelabai, -0.3
nuzudyti, -2	labai, 0.5
zudikas, -2	stipriai, 0.5