VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
PROGRAMŲ SISTEMŲ KATEDRA

E. portalų komentarų emocijų tyrybos metodika

Technique of emotion detection in comments of e. portals

Magistro baigiamasis darbas

Atliko: Klemensas Minkevičius (parašas)

Darbo vadovas: prof. habil. dr. Leonidas Sakalauskas (parašas)

Recenzentė: doc. dr. Audronė Lupeikienė (parašas)

Vilnius – 2017

SANTRAUKA

Svarbią šiuolaikinės žiniasklaidos dalį sudaro interneto portalai, kuriuose e. pavidalu talpinami straipsniai, o skaitytojai reiškia savo nuomones (komentarus), kurios gana dažnai būna anoniminės. Darbo autorius siekė tikslo patobulinti e. portalų komentarų tyrybos metodiką, leidžiančią šiuos komentarus įvertinti emociniu požiūriu.

Siekiant šio tikslo, buvo analizuojama prof. Bing Liu emocijų aptikimo tekste metodika, nustatant metodus, tinkamus e. portalų komentarų emocijų tyrybai, be to, kuriama emocijų aptikimo e. portalų komentaruose metodika, paremta žodyniniu ir Naive Bayeso metodais, ir ji buvo palyginama su Bing Liu metodika.

Darbo metu buvo išanalizuoti emocijų aptikimo tekste metodai nustatant metodus, tinkamus anoniminių e. portalų komentarų emocijų tyrybai; sukurtas patobulintos e. portalų komentarų tyrybos metodikos prototipas; sukurta klasifikatoriaus mokymo bazė; patobulinta prof. Bing Liu emocijų tyrybos metodika pagal šiuolaikinę komentarų analizės praktiką pridėtas mokymo duomenų paruošimo komponentas su duomenų paruošimo metodais ir nuo šiol į klasifikavimo įeina ir Naive Bayes'o klasifikatorius, ir leksikono metodas.

Pagal gautus rezultatus autorius formuoja tokias išvadas, kad, atlikus eksperimentą, įvertintas tikslumas, kuris viršija 67,5 proc. Be to, ši metodika gali būti taikoma įvairių temų komentarų įvertinimui emociniu požiūriu sudarius klasifikatoriaus mokymo duomenų bazę, paremtą ekspertinėmis išvadomis.

SUMMARY

Important part of nowadays‘ media consists of news portals, which have new articles daily. Readers of these articles express opinions (comments) which usually are anonymous. Author had purpose to upgrade technique of comments‘ mining, which lets to evaluate these comments by emotional aspect.

By seeking this purpose there had been analysed a technique of emotion detection in news portals‘ comments, which is based on lexicon and Naive Bayes methods, and also it was compared with technique of prof. Bing Liu.

The research has its results: methods of emotional detection were analysed by deciding which of them fit for mining of comments of news portals. At second, there were prototype of upgraded technique and learning base created. Futhermore, technique of prof. Bing Liu was upgraded by adding data preprocessing methods and changing methods of classifier component.

The resarch has its conclusions: after experiment author got 67,5 percent accuracy. Moreover, this technique can be adapted to various themes by creating classifier learning base, which was based on expert conclusions.

TURINYS

ĮVADAS

Aktualumas ir naujumas

Svarbią šiuolaikinės žiniasklaidos dalį sudaro interneto portalai, kuriuose e. pavidalu talpinami straipsniai, o skaitytojai reiškia savo nuomones (komentarus), kurios gana dažnai būna anoniminės. Šių komentarų turinys talpina daug informacijos, kurios analizė yra aktuali socialiniu ir verslo požiūriu. Tačiau tokia analizė dažniausiai atliekama rankiniu būdu, nes kompiuterinių anoniminių priemonių analizės metodai nėra plačiai naudojami, nes paprasčiausiai tokios priemonės nėra sukurtos. Darbe nagrinėjamas e. portalų komentarų tyrybos įrankio prototipas, taikant sentimentų analizės metodus.

Sentimentų analizė - pažangi duomenų tyrybos disciplina, susijusi su e. komentarų tyryba. E. portalų komentarų tyrybos metodika gali padėti automatizuoti komentarų e. portaluose vertinimą emociniu požiūriu. Sentimentų analizės tyrimų lauke sukurta ne viena emocijų aptikimo metodika [Mic2013] [Cha2012] [Mor2012]. Viena iš jų - sentimentų analizės autoriteto prof. Bing Liu metodika, kurią galima patobulinti pgl. dabartinius mokslinius tyrimus.

Darbo tikslas

Patobulinti e. portalų komentarų tyrybos metodiką, leidžiančią šiuos komentarus įvertinti emociniu požiūriu.

Uždaviniai

Siekiant šio tikslo, yra sprendžiami tokie uždaviniai:

- Išanalizuoti prof. Bing Liu emocijų aptikimo tekste metodiką, nustatant metodus, tinkamus anoniminių e. portalų komentarų emocijų tyrybai.
- Sukurti emocijų aptikimo e. portalų komentaruose metodiką, paremtą žodyniniu ir Naive Bayeso metodais, ir palyginti ją su Bing Liu metodika.

1. EMOCIJŲ TEKSTE APTIKIMO METODŲ APŽVALGA

1.1. Sentimentų analizės sinonimai ir apibrėžimas

Sentimentų analizė (Sentiment Analysis) – duomenų tyrybos subdisciplina [Dob2001], kuri turi ne vieną sinonimą: Nuomonių gavyba (Opinion Mining; Opinion Extraction), Subjektyvumo analizė (Subjectivity analysis), Sentimentų gavyba (Sentiment mining) [Jur2014]. Šioje tyrimų srityje susiejami nuomonė, sentimentai ir subjektyvumas. Tačiau kodėl nuomonė, atrodytų, racionalus ir logiškas dalykas, susiejamas su sentimentais ir subjektyvumu, kurie yra iracionalūs?

Pirmiausia, Websterio anglų kalbos žodyne [Web2017] *subjektyvus* (subjective) yra sietinas su tuo, kas paremta ne faktais, o jausmais ir nuomone (*based on feelings or opinions rather than facts*). Taigi anglų kalboje nuomonė (opinion) yra skiriama nuo faktiško – pagrįsto požiūrio. Sentimentai tame pačiame žodyne yra siejami su nepagrįstu *opinion* ir su subjektyviu *feeling*: sentimentas – tai požiūris, mintis ir vertinimas, kurį sužadino jausmas (*an attitude, thought, or judgment prompted by feeling*). Taigi sinonimiškai vartojamos sąvokos Sentimentų analizė, Sentimentų gavyba, Nuomonių gavyba ir Subjektyvumo analizė, atsiradusios angliakalbiame kontekste, iš tiesų laikytinos sinonimiškomis.

Sentimentų analizė galėtų būti apibrėžiama kaip duomenų tyrybos subdisciplina, kuri, naudodama duomenų tyrybos, kompiuterinės lingvistikos, statistikos ir mašininio mokymo metodus, tiria elektroninius tekstus ir iš jų išgauna nuomones (opinion), požiūrius (view; attitude) ir emocijas (emotions) [41]. Šio darbo autorius pasirinko Sentimentų analizės tyrimų sritį – Emocijų aptikimą (Emotion Detection), taigi darbo autorius koncentruojasi į emocijų tyrybą.

1.2. Sentimentų analizė: Tekstų tipai

Sentimentų analizė apima tekstus, kurie gali būti įvairūs. Jie galėtų būti klasifikuojami formalumo požiūriu ir teksto ilgio požiūriu.

1.2.1. Tekstai formalumo požiūriu

Formalūs tekstai pasižymi apibrėžtomis ir vienareikšmiškomis sąvokomis. Tai įvairūs dokumentai (pavyzdžiui, teisės aktai; taisyklės) ir vieši informaciniai pranešimai (pavyzdžiui, televizijos naujienų rašytiniai pranešimai).

Neformalūs tekstai pasižymi neformalia kalba, kuri neretai netaisyklinga, emocinga ar žargoniška – slengas, o slengas ir netaisyklinga kalba – iššūkis sentimentų analizės tyrėjams [40]. Taigi sentimentų analizės tekstai yra neformalūs.

1.2.2. Tekstai ilgio požiūriu

a) Sakinio lygis (tekstas = 1 sakinys).
b) Dokumento (tekstas>1 sakinio) lygis.

Darbo autorius pasirinko tyrinėti e. portalų komentarus, rašomus po e. portalų straipsniais. Tai neformalūs tekstai, kurių apimtis neretai varijuoja nuo sakinio lygio iki dokumento lygio (10 sakinių). E. portalų komentarai gali būti ir ilgesni, tačiau tuo atveju sudėtingėja jų analizė.

1.3. Sentimentų analizė: Lietuvių kalbos specifika

Sentimentų analizei svarbi lietuvių kalbos specifika, nes ši specifika diktuoja sentimentų analizės metodiką. Lietuvių kalbai būdingos dinamiškos morfologinės ypatybės, o dėl šių ypatybių sentimentų analizės metodika privalo turėti duomenų paruošimo metodus.

1.3.1. Vardažodžiai

Lietuvių kalboje vardažodžiai turi linksnio, giminės ir skaičiaus gramatines kategorijas.

1.3.1.1. Linksniai

Lietuvių kalboje yra šie septyni linksniai: vardininkas (nominatyvas), kilmininkas (genityvas), naudininkas (datyvas), galininkas (akuzatyvas), įnagininkas (instrumentalis), vietininkas (lokatyvas), šauksmininkas (vokatyvas).

Daiktavardžiai turi penkias linksniuotes, skirstomas į dvylika linksniavimo paradigmų. Paradigmą lengviausia nustatyti pagal daugiskaitos naudininko linksnio kamiengalio balsį, t. y. balsį, kuris eina prieš priebalsius. Pvz., darbams, sūnums, jūroms ir t. t. Kiekviena paradigma pasižymi tik jai būdingomis linksniavimo savybėmis.

1.3.1.2. Giminės

Vardažodžių giminės – vyriškoji ir moteriškoji. Lietuvių kalboje nėra bevardės giminės daiktavardžių. Būdvardžiai ir linksniuojamos veiksmažodžio formos lietuvių kalboje taip pat turi ir

bevardę giminę (tylu, matoma, ištirpę), kuri vartojama derinant su tam tikrais neapibrėžtais įvardžiais (kažkas, viskas, tai ir pan.).

1.3.1.3. Skaičius

Lietuvių kalboje dažniausiai sutinkamos vienaskaita ir daugiskaita, tačiau labai retai sutinkama ir dviskaita – trečia skaičiaus kategorija.

1.3.1.4. Įvardžiuotinės formos

Minėtina dar viena išskirtinė lietuvių kalbos ypatybė – įvardžiuotinės formos, kurios ryškiausios būdvardžiuose (taip pat jas turi kelintiniai skaitvardžiai, kai kurie įvardžiai ir dalyviai). Įvardžiuotinės formos nepažįstamos anglų, vokiečių, prancūzų ir šimtams kitų kalbų (bet jas turi rusų kalba). Įvardžiuotinės formos vartojamos pabrėžiant, išskiriant daiktą kartu su jo savybe iš daugelio tokių pat.

1.3.1.5. Laipsniavimas

Būdvardžiai ir prieveiksmiai, kaip ir daugelyje kalbų, laipsniuojami. Taip pat laipsniuojami ir dalyviai. Būdvardžiai ir prieveiksmiai lietuvių kalboje laipsniuojami iš esmės tik sintetiškai. Laipsnių skaičius 3-5: 3 pagrindiniai (nelyginamasis, aukštesnysis, aukščiausiasis) ir 2 tarpiniai (aukštėlesnis ir visų/pats aukščiausiasis). 4 iš jų reiškiami sintetinėmis ir tik visų (pats) aukščiausiasis – analitinėmis formomis.

1.3.2. Veiksmažodžiai

Lietuvių kalboje veiksmažodis turi daug neasmenuojamųjų formų: dalyvius, pusdalyvius, padalyvius, būdinius, siekinius. Bene sudėtingiausia iš jų dalyviai, turintys būdvardžių ir veiksmažodžių savybių. Dalyviai kaitomi skaičiais, linksniais, giminėmis, laipsniais, laikais, rūšimis.

1.4. Sentimentų analizė: Tekstų temos/sritys

Sentimentų analizės srityje populiarios šios tekstų temos ir sritys [Mul2013]:

- Šiuolaikinė politika ir politikos mokslai.
- Teisė ir įstatymai.
- Sociologija ir viešoji nuomonė.
- Psichologija ir asmens psichologinė būsena.

- Ekonomika ir rinkos tendencijos.

1.5. Sentimentų analizė: Populiariausios tyrimų sritys

Pgl. atliktų tyrimų skaičių yra išskiriamos trys populiariausios sentimentų analizės tyrimų sritys [Med2014]:

- Emocijų aptikimas (Emotion Detection) tekste.
- Resursų kūrimas (Building Resources), t.y. sentimentų žodynų ir tekstynų kūrimas.
- Mašininio mokymo šaltinių jungimas (Transfer Learning), t.y. sritis, kurioje siekiama apjungti skirtingų šaltinių ir įvairiakalbes (pavyzdžiui, anglų, lietuvių ir kinų) ontologijas. Ontologijų sąryšio (mapping) nebuvimas yra kliūtis panaudoti ontologijas sentimentų analizei.

1.6. Emocijų aptikimas tekste: Emocijų modeliai

Prieš emocijų aptikimo metodų analizę yra svarbu akcentuoti, kad yra ne vienas emocijų klasifikavimo modelis, bet trys modeliai yra svarbiausi ir dažniausiai naudojami Sentimentų Analizėje: Binarinis modelis, Emocijų kategorijų ir Emocijų Dimensijų modeliai.

1.6.1. Binarinis modelis

Binarinis modelis yra dažniausiai sutinkamas šiuolaikinėje sentimentų analizės praktikoje. Šio modelio tikslumas pakankamai didelis, nes emocijų klasės yra tik dvi.

1.6.2. Emocijų kategorijų modeliai

Emocijų kategorijų modelis yra skirtas apibrėžti emocijų klases [Can2014]. Kategorijų modeliu daroma prielaida, kad yra diskrečios emocijų kategorijos.

Pauliaus Ekmano (Paul Ekman) modeliu apibendrintos šios šešios pagrindinės emocijos: pyktis, pasišlykštėjimas, baimė, laimė, liūdesys ir nustebimas. Robertas Plučikas [Bin2010] apibrėžia emocijų modelį, kuris susideda iš šešių Ekmano emocijų ir dviejų Plučiko prijungtų emocijų – pasitikėjimo ir nekantrumo (anticipation). Šios aštuonios emocijos yra sujungtos į šiuos keturis antonimų rinkinius: džiaugsmas-liūdesys, pyktis-baimė, pasitikėjimas-pasišlykštėjimas, nustebimas-nekantrumas (anticipation).

1.6.3. Emocijų dimensijų modeliai

Emocijų dimensijų modeliai pristato emocijas dimensijų forma. Kiekviena emocija užima vietą erdvėje. Vienas iš emocijų dimensijų modelių yra Raselo afektu apskritimo modelis (Russell's Circumplex Model of Affect – RCMA) teigia [Des2013], kad emocijos gali būti išdėstytos dviejų dimensijų apskritimo erdvėje – vertikalioje ir horizontalioje (**Error! Reference source not found.** pav.). Horizontali dimensija apibrėžia, kiek emocija yra maloni (pleasant) ir kiek nemaloni (unpleasant). Vertikali dimensija apibrėžia, kiek emocija aktyvuojanti (activation) ir kiek emocija slopinanti (deactivation).

1 pav. Raselo afektų apskritimo modelis

Kitas emocijų dimensijų modelis yra Mehrabian'o PAD būsenų modelis [Des2013]. Modelis skirsto emocijas į malonumo-nemalonumo, aktyvumo-pasyvumo ir valdymo-paklusimo emocijas. Valdymo-paklusimo dimensija apibrėžia, kiek subjektas linkęs valdyti situaciją, o kiek paklusti.

Darbo autorius iš šių trijų emocijų pasirinko binarinį modelį dėl jo tikslumo. Be to, dar nėra sukurta aiškių modelių kriterijų, kaip atskirti vieną emocijos klasę nuo kitos, pavyzdžiui, įtampą nuo nuobodulio.

1.7. Emocijų aptikimas tekste: Metodų grupės

Emocijų aptikimo straipsniuose yra išskiriamos šios trys emocijų aptikimo metodų grupės:

- žodyniniai metodai (Lexicon-based approaches [Can2014]);

- mašininio mokymo metodai (Machine-learning approaches [Can2014] [Des2013] [Bin2010]);
- hibridiniai metodai (hybrid approaches [Des2013] [Bin2010]).

1.8. Emocijų aptikimas tekste: Žodyniniai metodai

Žodyniniai metodai yra metodai, kuriais naudojamas vienas ar keli leksiniai šaltiniai aptikti emocijas.

1.8.1. Raktažodiniai metodai

Tarp šių metodų galima rasti metodų, kurie klasifikuoja tekstą į emocijų kategorijas [Can2014]. Šie metodai atlieka paprastą algoritmą, kuris tikrina, ar yra emocingų žodžių antraštėse, ir skaičiuoja šių emocingų žodžių dažnumą tekste. Šie metodai naudoja WordNet'o afektų žodyną.

1.8.2. Ontologiniai metodai

Taip pat tarp žodyninių metodų randami ontologiniai metodai [Can2014]. Ontologiniai metodai naudoja Emotinet'ą – emocijų aptikimo šaltinį, kuris grįstas bendromis žiniomis apie konceptus (common sense knowledge on concepts), jų sąveiką ir jų emocinį rezultatą – aptikti emocijas. EmotiNet'o modeliai veikia kaip veiksmų grandinės ir jų atitinkamas emocinis atsakas bei naudoja naudoja ontologinę reprezentaciją. Metodų vertinimas susideda iš testavimo, ar, įdarbinant modelį ir EmotiNet'o žinias, galima aptikti emocijas naujuose pavyzdžiuose. Metodų vertinimas rodo, kad EmotiNet'o struktūra ir turinys yra tinkami kreiptis į EmotiNet'o emocijų tinklą.

1.8.3. Statistiniai metodai

Statistiniai metodai taip pat yra laikomi žodyniniais [Can2014]. Dauguma žiniomis gristų (knowledge – based) darbų naudoja Latentinę Semantinę Analizę (Latent Semantic Analysis – LSA), statistinį metodą analizuoti ryšius tarp dokumentų rinkinio ir terminų, pavartotų šiuose dokumentuose, kad būtų sukurti prasmingi šablonai, susiję su dokumentais ir terminais. Alaistair Gill'as [Gil2005] panaudojo LSA ir Hypererdvės analogą kalbai (Hyperspace Analogue to Language – HAL) metodą, kad būtų paskaičiuotas semantinis panašumas tarp tekstų ir emocinių raktažodžių. Wang'as ir Zeng'as [Yam2015] pristatė metodą, kuris naudoja patobulintą LSA algoritmą teksto emocijų klasifikavimui pgl. ISEAR duomenų rinkinį.

1.8.4. Tekstyno metodai

Tyrinėtojai į savo eksperimentus įtraukia įvairius tipų tekstynus [Kim2011]. Wilson'as pasiūlė naują metodą analizuoti sakinio lygio sentimentų analizę. Priešingai, dauguma sentimentų analizės darbų buvo atliekama dokumento lygiu. Wilson'as, Wiebe'as ir Hoffmann'as [SRW2007] pradėjo naudoti dviejų žingsnių metodą, kuris klasifikuoja kiekvieną frazę kaip neutralią arba poliarišką ir tada diferencijuoja visas poliariškas frazes į pozityvias, negatyvias, prieštaringas (both) arba neutralias. Šis dviejų žingsnių klasifikatorius pavadintas Neutralus-Poliariškas klasifikatorius (Neutral-Polar Classification), kuris naudoja Daugiaperspektyvį klausimų atsakymų (Multi-perspective Question Answering – MPQA) nuomonių tekstyną. Anotuotos 15,991 subjektyvių tekstyno frazių iš 425 dokumentų (8,984 sakiniai). Klasifikatorius sukurtas naudojant BoosTexter AdaBoost.HM mašininio mokymo algorimą.

Nauji metodai pademonstruoti, kad būtų galima vertinti emocijas interneto komunikacinėse priemonėse, pavyzdžiui, AOL Instant Messenger ar MSN Messenger. Tekstinės pokalbių žinutės yra automatiškai konvertuojamas į šnekos fonemas, nes žmonės gali dažnai atpažinti vieni kitų emocijas jusdami balso toną nei skaitydami tekstą. Pirminis privalumas yra tas, kad metodas yra toks atsparus pokalbių (*angl.* chat) duomenų triukšmui dėl šnekos fonemų savybių. Tiksliau rašant, daugybė interneto pokalbių žinučių yra su klaidomis: nesilaikoma gramatikos taisyklių, žodžiai nepabaigiami. Tekstyno duomenys pgl. šį metodą yra paimami iš dviejų šaltinių. Pirmas pokalbių rinkinys yra iš dviejų mėnesių dviejų žmonių pokalbių. Antras pokalbių rinkinys yra kitų dviejų žmonių dialogai. Be to, k-arčiausias (k-nearest) Instance Based Learning (iBk) narys yra įdarbinamas taikyme ir palyginamas su kitais mašininio mokymo metodais (Naïve Bayes, One R, Decision Table, j48). Kita vertus, Litman'as and Riley'is naudoja akustines-prosodines ir leksines savybes, kad nuspėtų studentų emocijas kompiuterio-žmogaus šabloniniuose klausimas-atsakymas dialoguose.

Tekstynas susideda iš student pokalbių su ITSPOKE (Intelligent Tutoring SPOKEn dialogue system - 2004) ir buvo rinkti nuo 2003 m. lapkričio iki 2004 m. gegužės. Kol leksinės savybės lenkia akustines-prozodines savybes, leksinių ir šnekos savybių susiejimas nepatobulina emocijų atpažinimo proceso, kaip buvo indikuota mašininio mokymosi eksperimentuose (patobulinti sprendimų medžiai – boosted decision trees). Liu, Lieberman'as, and Selker'is sukuria galingą metodą, kuris leidžia naudoti galingą realiojo pasaulio bendrųjų prasmių duomenų bazę. Šis metodas sprendžia problemas ir naikina apribojimus, kuriuos keturios egzistuojančių metodų kategorijos (raktažodžių koregavimas, leksinis giminingumas, statistinis natūralios kalbos apdorojimas,

rankiniai metodai) turi. Open Mind Common Sense (OMCS) duomenų bazėje surinkta apie 400 000 faktų apie kasdienį gyvenimą. Bendrosios prasmės (CommonSense) yra susietos su anglų k. sakiniais, kurie yra padalyti į 20 ar daugiau šablonų. Keturi bendrųjų prasmių emocijų modeliai (Subjekto-Veiksmažodžio-Objekto modelį, Konceptualų unigramos modelį, Konceptualų valentingumo modelį) yra sudaryti OMCS ir yra sujungti su tikslu klasifikuoti emocijas tekste. Be to, keturi modeliai kaip irimas (decay), interpolation (interpoliacija), globalinė nuotaika (global mood) ir meta-emocijos yra taikomi, kad būtų įgalinamas emocijų perėjimas nuo vieno sakinio iki kito. Šiame taikyme Sistema integruota į kliento e. paštą, kad būtų teikiamas grįžtamasis ryšys vartotojui.

1.8.5. Tezauriniai metodai

Tezauriniai metodai naudoja sinonimų žodynų, retai vartojamų žodžių žodynus su tikslu apibrėžti emocijų poliariškumą žodžiuose, sakiniuose ir dokumentuose [Kim2011]. Yi, Nasukawa, Bunescu, and Niblack naudoja kompiuterinės lingvistinės metodus sentimentų analizėje, kas yra emocijų aptikimas ir subjekto bei sentimento asociacijų analizė. Du savybių algoritmai yra sukurti ir ištestuoti remiantis maišytos kalbos modeliu ir panašumo koeficientu. Sentimentų analizė antrame etape naudoja du lingvistinius resursus: sentimentų žodyną ir sentimentų šablonų duomenų bazę. Sentimentų žodynas susideda iš sentimentų apibrėžimų iš General Inquirer (GI), Emocingos kalbos žodyno (Dictionary of Affect of Language - DAL) ir WordNet'o. Sentimentų šablonų duomenų bazę sudaro sentimentų aptikimo šablonai sakinio predikatams, kuriame sentimentų veiksmažodžiai yra surinkti iš GI, DAL ir WordNet'o. Maždaug 120 sentimentų predikatų šablonų yra saugoma duomenų bazėje. Eksperimento rezultatas yra palyginamas su kolokacijos algoritmu ir ReviewSeer klasifikatoriaus algoritmu. Taisylėmis grįstu būsenų analizės modelis buvo sukurtas susidoroti ir su taisyklingu formaliu, ir su netaisyklingu neformaliu rašytiniu tekstu.

Nuo tada, kai neformalios žinutės e. erdvėje yra dažnai rašomos sutrumpinta ar ekspresyvia emocinga kalba, į šį faktą atsižvelgiama. Šis modelis naudoja 1627 būdvardžius, daiktavardžius, veiksmažodžius ir prieveiksmus, paimtus iš WordNet'o. Be to, 364 jaustukai ir 337 akronimai ir santrumpos yra surinkti, kad būtų susidorojama su santrumpų kalba ir vyktų sėkminga jaustukų, santrumpų ir kitų žodžių savybių interpretacija. Emocijų analizės modelis susideda iš penkių sekos žingsnių: simbolinė ženklų (cue) analizė, sintaksinė struktūros analizė, žodžio lygio analizė, frazės lygio analizė ir sakinio lygio analizė.

1.9. Emocijų aptikimas tekste: Žodyninių metodų kritika

1.9.1. Raktažodžių dviprasmiškumas

Nors emocingi raktažodžiai yra tiesus kelias aptikti emocijų asociacijas, raktažodžiai gali būti keliaprasmiški arba jų prasmės gali būti neaiškios. Iškirpus šiuos žodžius iš teksto, dauguma žodžių gali prarasti prasmes dėl kitos naudojimo srities ar naudojimo konteksto. Be to, net minimalus etikečių (label) rinkinys (be visų jų sinonimų) gali turėti skirtingas emocijas, tam tikrais atvejais tai gali būti ironiški arba ciniški sakiniai.

1.9.2. Negebėjimas atpažinti sakinių be raktažodžių

Raktažodiniai metodai yra visiškai pagrįsti emocijų raktažodžiais. Tačiau sakiniai be raktažodžių implikuoja, kad jie visiškai neemocingi, kas neabejotinai klaidinga. Pavyzdžiui, frazės „Aš šiandien išlaikiau egzaminą“ ir „Jėėė! Aš šiandien išlaikiau egzaminą“ turėtų sukelti džiaugsmo emociją, bet džiaugsmo emocija gali būti neaptikta, jei vienintelis raktažodis išreikšti džiaugsmą - „jėėė“.

1.9.3. Lingvistinės informacijos nepakankamumas

Sintaksinės struktūros ir semantika taip pat turi įtakos išreikštoms emocijoms. Pavyzdžiui, „Aš juokiausi iš jo“ ir „Jis juokėsi iš manęs“ galėtų būti laikomos skirtingomis emocijomis iš pirmo žmogaus perspektyvos. Taigi lingvistinės informacijos ignoravimas sukelia problemų raktažodiniams metodams. Raktažodiniai metodai turėtų ne tik aptikti egzistuojančius raktažodžius, tačiau ir turėtų aptikti lingvistinę informaciją, kad galėtų įvertinti emocijas daug tiksliau.

1.10. Emocijų aptikimas tekste: Mašininio mokymo metodai

Mašininis mokymas yra mokslinė disciplina, kuri tiria algoritmus, kurie parenka metodo parametrus. Mašininio mokymo algoritmai yra naudojami aptikti emocijas. Šie mašininio mokymo metodai skirstomi į prižiūrimus (supervised) ir neprižiūrimus (unsupervised) mašininio mokymo metodus.

1.10.1. Prižiūrimo mašininio mokymo metodai

Prižiūrimas mokymas susijęs su pažymėtais (labelled) mokymo duomenimis, mokymo rinkinių pavyzdžiais ir dėl to reikalauja žmogaus įsikišimo. Prižiūrimo mokymo algoritmas analizuoja mokymo duomenis ir nusprendžia, kurie mokymo duomenis skirti klasifikavimui [Can2014]. Pažymėtas tekstynas yra didelis ir struktūruotas tekstų rinkinys, kuris neretai anotuotas

emocijų tagais. Šiuo atveju anotacijos procesas yra laikomas vienu iš didžiausių jų trūkumų, nes jis tampa nuobodi ir laiką eikvojanti užduotis. Deja, jau yra atlikta darbų, susijusių su emocijų aptikimu Twitter‘io žinutėse, kur mokymo pavyzdžiai yra automatiškai pažymimi hashtag‘ais ir jaustukais tarp kitų [RP2013]. Straipsnyje siūloma automatizuoti šį mokymo duomenų žymėjimo procesą. Straipsnio autorius teigia, kad hashtag‘ai yra geri emocijų žymekliai. Apsvarsčius darbus, kurie taiko prižiūrimo mokymo algoritmus, galima rasti naudojamų ir kategorijų, ir dimensinių emocijų modelių emocijų aptikimui Kategoriniai emocijų modeliai yra dažniausiai naudojami emocijų aptikime. Vienas iš pirmųjų darbų, grįstas šiuo modeliu, yra aprašytas [ARS2005]. Šiame straipsnyje buvo pristatyta prižiūrimo mašininio mokymo taikymas su SnoW mokymo architektūra. Straipsnio autoriai anotavo tekstyną su išplėstu Ekmano pagrindinių emocijų rinkiniu. [SM2007] viename eksperimente taiko Naive Bayes klasifikatorių, kuris apmokytas su blogo antraštėmis iš LiveJournal.com. Straipsnyje naudotas rinkinys blogo žinučių, anotuotų su Ekmano emocijomis. [RMN2012] pristatė emocijų klasifikatorių, kuris gali apibrėžti rašančio žmogaus emocijų klasę. Straipsnio autoriaus klasifikatorius yra pagrįstas multiklasiu SVM kerneliu ir priima sprendimus remdamasis pagrindinėmis emocijomis, apibrėžtomis [Eck1999]. Roberts [RRJ2012] taip pat naudoja Ekmano šešias pagrindines emocijas, tačiau išplečia jas meilės emocija (LOVE).

Straipsnyje aprašyta sistema naudoja seriją binarinių SVM klasifikatorių aptikti kiekvieną iš septynių emocijų. Kitas su emocijų modeliais susijęs darbas Suttles ir Ide [SI2013] klasifikuoja emocijas remdamasis Plučiko (Plutchick) aštuonių polinių emocijų rinkiniu. Tai leidžia išspręsti emocijų klasifikavimo daugiaklasiškumo problemą. Ši sistema naudoja Nutolusį Prižiūrimą mokymą (Distant Supervision). Sistemos kūrėjai naudoja Raselo emocijų apskritimo modelį kaip emocijų modelį ir moko prižiūrimus klasifikatorius aptikti emocijas.

1.10.2. Neprižiūrimo mašininio mokymo metodai

Neprižiūrimo mokymo algoritmai skiriasi nuo prižiūrimo mokymo tuo, kad nereikalauja jokių pažymėtų duomenų ir žmogaus įsikišimo [Can2014]. [SM2007] taiko neprižiūrimo mašininio mokymą kombinuodami su LSA ir su WordNet‘o emocijų žodynu. Šiame sprendime naudojamas Ekmano pagrindinių emocijų modelis. Agrawal [Agr2012] siūlo Novel Neprižiūrimą Konteksto metodą, kuris nepriklauso nuo jokio egzistuojančio emocijų žodyno. Sprendimo modelis yra pakankamai lankstus, kad suklasifikuotų sakinius pgl Ekmano šešių pagrindinių emocijų modelį. Calvo and Kim [CK2013] pristato skirtingas kategorinį sprendimą, kuris paremtas Vektoriaus Erdvėje Modeliu (Vector Space Model – VSM) ir trimis dimensinėmis redukcijos technikomis:

Latentine Semantine Analize (LSA), Tikimybine Latentine Semantine Analize (TLSA) ir Nenegatyvia Matricos Faktorizacija (NMF).

1.11. Emocijų aptikimas tekste: Mašininio mokymo metodų kritika

1.11.1. Sunkumai apibrėžiant emocijų indikatorius

Pirmoji problema yra ta, kad nors mokymu grįsti metodai gali automatiškai apibrėžti tikimybes tarp savybių ir emocijų, mokymo metodams dar reikia raktažodžių, bet ne ta savybių forma. Pačios intuityviausios savybės gali būti jaustukai, kurie gali būti vertinti kaip autoriaus emocijų anotacijos tekste. Pakopų (cascading) problemos gali būti tokios pačios kaip tos, kurios yra raktažodiniuose metoduose.

1.11.2 Per paprastos emocijų kategorijos

Dauguma mokymo metodų gali tik klasifikuoti sakinius į dvi kategorijas, kurios yra teigiamos arba neigiamos. Nors emocijų etikečių (labels) skaičius priklauso nuo taikomo emocijų modelio, praktinėse sistemose būtų tikimasi apibrėžti daugiau kategorijų.

1.12. Emocijų aptikimas tekste: Hibridiniai metodai

Nuo tada, kai buvo suvokta, kad raktažodiniai metodai su tezauru ir Naive mokymo metodai negali pasiekti patenkinamų rezultatų, kai kurios sistemos naudoja hibridinius metodus kombinuodami ir žodyninius, ir mašininio mokymo metodus, kas leidžia patobulinti tikslumą ir apibrėžti kategorijas [Des2013] [Bin2010]. Pati reikšmingiausia hibridinė sistema yra Wu, Chuang ir Lin darbas, kuris naudoja taisyklėmis grįstą priėjimą prie semantikos, susijusios su specifinėmis emocijomis ir kinų kalbos ontologijomis, kad būtų išskirti atributai.

1.12.1. Sintezės metodas

Sintezės (fusion) metodas yra hibridinis metodas, kuris naudoja tekstyno ir tezauro metodus, kad įveiktų vienas kito trūkumus. [Kim2011] demonstruoja hibridinę sistemą, kuri yra išskaidyta į raktažodžių ir mašininio mokymo metodus. Jei įvesties sakinys turi emocingų raktažodžių, tuomet naudojami raktažodiniai metodai. Kitais atvejais sistema naudoja mašininio mokymo metodus. Raktažodinis metodas pagrįstas EKD (Emocingų raktažodžių žodynu - Emotional Keyword Dictionary), kuris susideda iš žodžių, kurie turi emocinį turinį. Kita vertus, KBANN tinklas naudoja 3200 sakinių, emociškai pažymėtų tekstyne, kurie surinkti iš dramos pjesių, romanų ir blog‘ų. Šita

emocijų atpažinimo sistema jungia raktažodinį metodą ir mašininio mokymo metodą ir išskiria šias aštuonias emocijų klases: pyktį, baimę, viltį, liūdesį, meilę, padėką, neutralumą kaip atskiras klases. Somasundaran‘o, Ruppenhofer‘io and Wiebe’o [SRW2007] mokslinis darbas teigia, kad klasifikatoriai su leksinėmis ir diskurso žiniomis našiausi aptikti emocijas grupiniuose pokalbiuose. Sentimentai ir ginčai (arguing) yra dvi anotacijos kategorijos. Šiame darbe 7 scenarijais grįsti komandų susitikimai (6504 sakiniai) iš AMI tekstyno [Car2006], kur dalyviai turėjo suprojektuoti naują TV valdymo pultą, yra anotuoti. Be to, Somasundran‘as, Ruppenhofer‘is ir Wiebe‘as naudoja ne tik sentimentų ir ginčų leksikonus kaip žinių šaltinius, bet taip pat dialogų aktus (Dialog Acts – DA) ir gretimumo poras (Adjacency Pair – AP) užfiksuoti diskurso vyksmą. Autoriai iškelia hipotezę, kad žinių šaltiniai, DA ir AP yra naudingi emocijų raiškos identifikatoriai pokalbių duomenyse.

Eksperimentai, kuriuos atliko Strapparav‘as ir Mihalcea [SM2007], naudojami palyginti tezaurinių ir tekstyno metodus su šešiomis emocijomis: pykčiu, pasišlykštėjimu, pykčiu, džiaugsmu, liūdesiu ir nuostaba. Tiksliau sakant, penkios skirtingos sistemos yra įdiegtos, kurios naudoja šiuos du metodus. *WordNet-Affect Presence, LSA Single Word, LSA Emotion Synset* and *LSA All Emotion Words* yra pagrįsti tezauriniu metodu ir Naive Bayes, apmokytu blog‘ų informacija, yra paskirtas tekstynų lingvistikos metodui. Kol žodyninis emocijų anotavimas yra svarbus, WordNet-Affect naudojama kaip emocingų žodžių duomenų baze. Priešingai blog‘ų antraštės iš LiveJournal.com yra tekstyno emocijų anotavimui. Be to, Latentinė semantinė Analizė (LSA) yra diegiama, kad pademonstruotų žodyninio metodo žodžių rinkinius ir tekstus, kai Naive Bayes klasifikatorius yra mokomas tekstyno metodo blog‘ų žinutėmis.

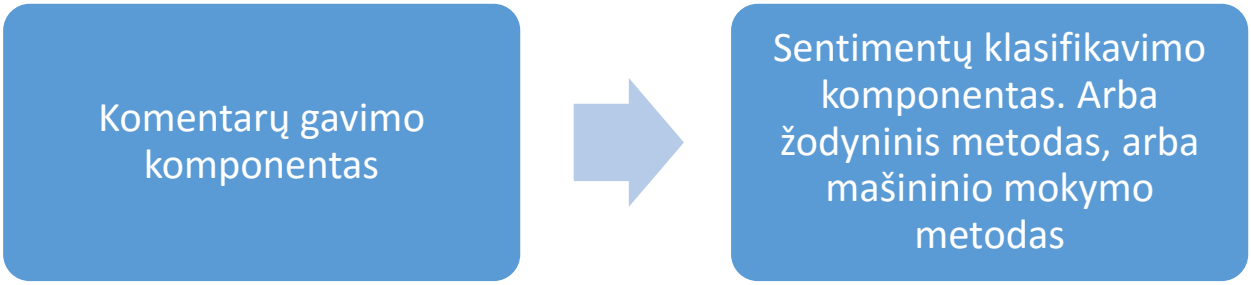
1.13. Emocijų aptikimas tekste: Hibridinio metodo kritika

1.13.1. Metodų suderinamumui reikia gilaus metodų išmanymo.

Hibridinis metodas, kaip metodų junginys, reikalauja gilios jungiamų metodų analizės. Tobulas hibridinis metodas yra tas, kai metodai ne tik išreiškia savo privalumus, tačiau ir padengia vienas kito trūkumus.

2. PROF. BING LIU KOMENTARŲ ANALIZĖS METODIKA

Prof. Bing Liu teigia [Liu2011], kad sentimentų analizė paprastai laikoma dviejų etapų procesu (2 pav.). Pirmo etapo metu dokumentai yra sugrupuojami pgl. temą (topic), ką daro tradicinės informacijos aptikimo ar paieškos sistemos. Sentimentų įverčiai gali būti gaunami naudojant arba mašininiu mokymu grįstą klasifikatorių (pavyzdžiui, Bayes'o Multinomial klasifikatorių - BNM ar Support Vector Machine - SVM), arba žodyniniu metodu grįstą sentimentų žodyną ir įverčių apdorojimo funkciją, kuri jungia sentimentų įverčius ir sentimentus.



2 pav. Prof. Bing Liu komentarų analizės metodika

2.1. Aptikimo komponentas

Komponentas atlieka tradicinį informacijos aptikimo (*angl*. information retrieval) uždavinį. Jis naudoja raktažodžius ir konceptus. Konceptai yra vardinės esybės (pvz., žmonių vardai-pavardės arba įvairių tipų frazės iš žodynų arba kitų šaltinių (pvz., Wikipedia). Užklausos veikimo strategija tokia: algoritmas iš pradžių atpažįsta and atskiria konceptus vartotojo užklausoje ir tada jų ieško e. portalų komentaruose.

2.2. Sentimentų klasifikavimo komponentas

Šis komponentas atlieka du uždavinius: pirma, klasifikuoja kiekvieną dokumentą į dvi kategorijas – sentimentalizuotą ir nesentimentalizuotą tekstus; antra, klasifikuoja kiekvieną sentimentalizuotą komentarą kaip tą, kuris išreiškia pozityvias, negatyvias ar mišrias emocijas. Abiems uždaviniams metodika naudoja prižiūrimą mokymą. Pirmame uždavinyje metodikos prototipas gauna didelį kiekį informacijos iš www.rateitall.com ir www.opinions.com. Duomenys yra taip pat surinkti iš temų, kurios yra vartotojų prekės ir paslaugos, valdžios suteikiamos teisės ir politiniai požiūriai. Nesentimentalizuoti mokymų duomenys yra gaunami iš svetainių, kurios teikia objektyvią informaciją (kaip antai, Wikipedia). Iš šių mokymo duomenų yra sukuriamas SVM klasifikatorius.

Šis klasifikatorius yra taikomas kiekvienam aptiktam pgl. temą komentarui. SVM klasifikatorius tada klasifikuoja kiekvieną komentarą kaip sentimentalizuotą ir nesentimentalizuotą. Jei komentaras priskiriamas sentimentalizuotų komentarų klasei, tada jo stiprumas (“its strength”), taip pat nustatomas. Komentaras laikomas sentimentalizuotu, jei bent vienas jo žodis sentimentalizuotas.

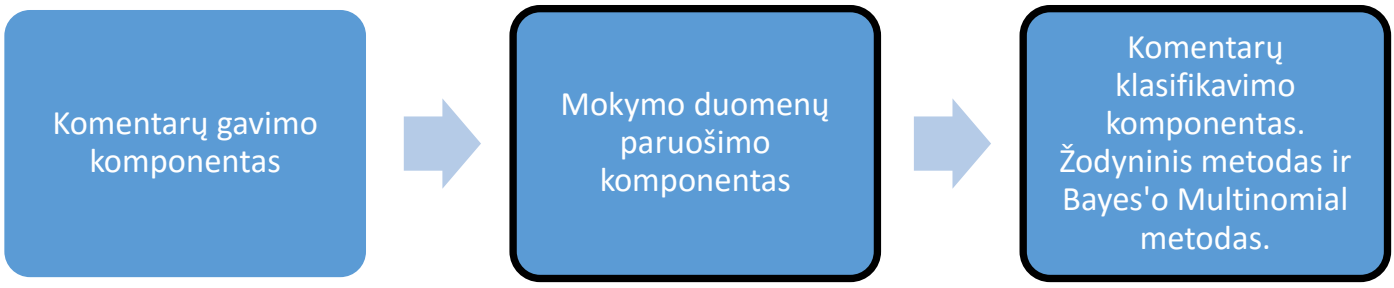
Norint apibrėžti, kuris komentaras išreiškia pozityvią, negatyvią ar mišrias emocijas, antrasis klasifikatorius yra sukonstruotas naudojant apžvalgas iš apžvalgų svetainių; komentarai tampa mokymo duomenimis. Žemas įvertis nurodo į neigiamą sentimentą ir priešingai – aukštas įvertis rodo į pozityvų sentimentą. Naudojant pozityvius ir negatyvius komentarus kaip mokymo duomenis, sentimentų klasifikatorius yra sukurtas klasifikuoti kiekvieną komentarą kaip turintį pozityvių, negatyvių ar mišrių sentimentų.

2.3. Bing Liu metodikos metodikos trūkumai

- Metodika neatsižvelgia į mokymo duomenų savybių normalizavimą.
- Metodika neatsižvelgia į mokymo duomenų apimties normalizavimą.
- Metodika nerekomenduoja nei vieno konkretaus mašininio mokymo algoritmo, kuris tiksliausias.

3. PATOBULINTOS METODIKOS TEORINIS PAGRINDIMAS

Patobulinta komentarų analizės metodika atsižvelgia ir į mokymo duomenų paruošimą, ir į klasifikavimo komponento optimizavimą (3 pav.).



3 pav. Patobulinta prof. Bing Liu komentarų analizės metodika

3.1. Komentarų gavimo komponentas

Straipsnių komentarai metodikoje gaunami iš semantiškai susijusių straipsnių. Tai daroma siekiant komentarų semantinio vientisumo.

3.2. Mokymo duomenų paruošimo komponentas

Mokymo duomenų paruošimas – tai operacijos, kurios skirtos sukurti duomenis, kurie skirti apmokyti emocijų klasifikatorių. Apačioje pateikiamos šios operacijos.

3.2.1. Skyrybos ženklų panaikinimas ir raidžių vertimas mažosiomis

Pirmoji operacija skirta pašalinti skyrybos ženklus, nes jie nereikalingi emocijų atpažinimui. Raidžių vertimas mažosiomis leidžia normalizuoti tekstą.

3.2.2. Jausmukų normalizacija

Jausmukai yra simboliai (4 pav.), kurie dažnai rašomi interneto komentaruose. Ši operacija sumažina jausmukų skaičių iki dviejų klasių - šypsena_pozityvi ir šypsena_negatyvi.

Icon											Emoji	Meaning
:-) :)	:-] :]	:-3 :3	:->> :>	8-) 8)	:-} :}	:o)	:c)	:^)	=]	=)	☺😊😊😀😀	Smiley or happy face.[4][5][6]
:-D :D	8-D 8D	x-D xD	X-D XD	=D	=3	B^D					😎😎😎😍	Laughing,[4] big grin,[5][6] laugh with glasses[7]
:-))												Very happy or double chin[7]
:-(:(	:-c :c	:-< :<	:-[:[	:-||	>:[	:{	:@	>:(			😞😞😭😞😦😞😞😞	Frown,[4][5][6] sad,[8] angry,[7] pouting
:'-(:'(											😢😭	Crying[8]
:'-) :')											😊	Tears of happiness[8]
D-':	D:<	D:	D8	D;	D=	DX					😨😧😫😱😩😕	Horror, disgust, sadness, great dismay[5][6] (right to left)

4 pav. Jausmukų pavyzdžiai

3.2.3. Keiksmažodžių ir žargono normalizacija

Ši operacija leidžia pakeisti slengą su jo formalia reikšme (l8 -> late), naudojantis sąrašu. Tai taip pat leidžia pakeisti keiksmažodžius ir užgaulius žodžius tagu “blogas_žodis”. Šių žodžių naudojimo motyvacija yra sumažinti triukšmą tekste ir patobulinti klasifikatoriaus veiklą.

3.2.4. Perteklinių žodžių panaikinimas

Pertekliniai žodžiai yra žodžiai, kurie neneša jokio emocinio krūvio: jungtukai, įvardžiai ir t.t. Tai yra svarbu šių žodžių išvengti mokyme, nes jie gali mažinti klasifikacijos tikslumą. Šie žodžiai paprastai yra neutralaus emocingumo. Pašalinus neutralius žodžius lieka dvi tikimybės – arba žodis priklauso teigiamai klasei, arba žodis priklauso neigiamai klasei.

3.2.5. Kamieninimas

Kamieninimo operacija leidžia rasti žodžio variacijas (daiktavardžius, būdvardžius ir prieveiksmius) ir jas priskirti vienam kamienui taip palengvinant mokymą ir klasifikavimą. Žodžio variacijos, kaip antai, ,,geras”, ,,geresnis”, ,,geriausias” ir ,,gerai” tampa kamienu ,,ger”.

3.2.6. Žodžio kamieno pozityvios emocijos tikimybės nustatymas

Nors kamieno emocingumo nustatymas yra subjektyvus procesas, tačiau kamieno emocingumo nustatymo fenomenologinis objektyvumas įmanomas, kai kamienų emocingumą vertina grupė. Magistro darbo autoriui pavyko suorganizuoti dviejų vertintojų grupę.

Kamienų emocingumo nustatymui sukurtos penkios emocijų klasės – dvi teigiamos, dvi neigiamos ir neutrali (1 lentelė). Jei pozityvaus kamieno tikimybė > 0,5, tai, tikėtina, yra pozityvus

kamienas. Jei pozityvaus kamieno tikimybė < 0,5, tikėtina, tai yra negatyvus kamienas. Jei pozityvaus kamieno tikimybė = 0,5, tikėtina, tai yra neutralus kamienas.

1 lentelė. Kamienų emocijų klasės ir jų pozityvaus kamieno tikimybės

Emocijos klasė	Pozityvaus kamieno tikimybė
Labai pozityvi	1
Pozityvi	0,75
Neutrali	0,5
Negatyvi	0,25
Labai negatyvi	0

3.3. Komentarų klasifikavimo komponentas

Kiekvieno žodžio prasmė su savimi „neša“ emocinį krūvį – teigiamą, neigiamą arba neutralų. Kuo frazėje teigiamų ar neigiamų žodžių, tuo frazė yra emocingesnė.

Norint suklasifikuoti komentarus į teigiamus ir neigiamus, atliekami šie žingsniai.

3.3.1. Komentaro P(T) skaičiavimas

Pagal Naive Bayeso multinomial klasifikatorių suskaičiuojama komentaro tikimybė priklausyti teigiamai klasei [5 pav.].

$$\eta = \sum_{i=1}^{N} [\ln(1 - p(T)_i) - ln(p(T)_i)]$$

P(T) – tikimybė, kad komentaras – teigiamas.

p(T) – ekspertų išvadomis gaunama tikimybė, kad kamienas teigiamas.

$$P(T) = \frac{1}{1 + e^{\eta}}$$

5 pav. Neigiamo komentaro tikimybės skaičiavimo formulė

3.3.2. Komentaro P(N) skačiavimas

Pagal Naive Bayeso multinomial klasifikatorių suskaičiuojama komentaro tikimybė priklausyti neigiamai klasei [6 pav.].

$$\eta = \sum_{i=1}^{N}[\ln(1 - p(N)_i) - ln(p(N)_i)]$$

P(N) – tikimybė, kad komentaras – neigiamas.

p(N) – ekspertų išvadomis gaunama tikimybė, kad kamienas neigiamas

$$P(N) = \frac{1}{1 + e^{\eta}}$$

6 pav. Teigiamo komentaro tikimybės skaičiavimo formulė

3.3.3. Tikimybių palyginimas

Atliekami tikimybių palyginimas ir komentaro emocingumo skaičiavimas

Pgl. Bajeso teoremą - jei p(T) > p(N), tada komentaras teigiamas.

Priešingai - jei p(N) > p(T), tada komentaras neigiamas.

Jei p(T) = p(N), tai arba komentaras nėra nei neigiamas, nei teigiamas, arba neutralus.

3.4. Bayeso klasifikatoriaus taikymo metodikoje pagrindimas

Svarbus žingsnis apmokant klasifikatorių – nuspręsti, kokį mokymo algoritmą naudoti. Pagal sentimentų analizės praktiką metodikose dažniausiai sutinkami šie metodai:

- Atraminis vektorių klasifikavimo algoritmas (angl. *Support Vector Machine – SVM*);
- Bajeso naivusis algoritmas (angl. *Naïve Bayes*);

Siūlomoje metodikoje naudojamas Naïve Bayeso metodas analogiškas taikomam spamo filtravimo algoritmuose.

Šio algoritmo pasirinkimą lėmė tai, kad jis pagal šiuolaikinę sentimentų analizės praktiką gana puikiai klasifikuoja tekstinę informaciją ir jo įgyvendinimas nėra labai sudėtingas. Tačiau, norint pasiekti dar geresnius klasifikavimo rezultatus, algoritmai reikalauja didelio mokymo duomenų kiekio. Kuo didesnis mokymo duomenų kiekis, tuo klasifikatorius tikslesnis.

3.5. Metodikos rezultatų patikrinimas

Prieš metodikos naudojimą reikia ją ištestuoti, t.y. patikrinti jos tikslumą (7 pav.). Testavimo proceso metu yra įvertinamos klasifikatoriaus priskirtos klasės, jas analizuojant, su testavimo duomenų klasėmis. Tikslumas, tai ta procentinė atvejų dalis, kai priskirtos klasės sutampa su klasifikavimo duomenų klasėmis (2 lentelė). Jis randamas taip:

$$Tikslumas = \frac{teisingai\ suklasifikuoti\ duomenys}{visas\ testavimo\ duomenų\ kiekis} = \frac{tt+tn}{nt+nn+tt+tn}$$

7 pav. Metodikos tikslumo skaičavimo formulė

Teisingai suklasifikuoti duomenys reiškia, jog išmokytas klasifikatorius nuspėja tą pačią klasę kaip ir klasę, apibrėžtą testavimo duomenyse.

2 lentelė. Klaidų matrica

		Spėta klasė	
		Teisingai	Neteisingai
Žinoma klasė	Teisingai	teisingai teigiamas (tt)	neteisingai teigiamas (nt)
	Neteisingai	neteisingai neigiamas (nn)	teisingai neigiamas (tn)

Teisingai teigiamas (tt) – teisingai priskirtų klasių skaičius, kai klasė yra teigiama.

Teisingai neigiamas (tn) – teisingai priskirtų klasių skaičius, kai klasė yra neigiama.

Neteisingai neigiamas (nn) – neteisingai priskirtų klasių skaičius, kai klasė yra neigiama.

Neteisingai teigiamas (nt) – neteisingai priskirtų klasių skaičius, kai klasė yra teigiama.

3.6. Patobulintos metodikos pranašumo tikslumu ir mokymo aibės apimtimi įrodymas pavyzdžiu

Šiame skyrelyje pateikiamas teorinis pagrindimas, kad patobulinta metodika yra tikslesnė nei prof. Bing Liu metodika ir kad patobulinta metodika naudoja mažesnę mokymo aibę. Teoriniam

pagrindimui naudojamas komentaro pavyzdys: „Tad tas namas buvo gražus ir jaukus, bet mažas. :(“.

3.6.1. Bing Liu metodikos žingsniai

- Iš mokymo bazės gaunamos tikimybės būti pozityviems

> **Rezultatas:** Tad – 0,5, tas – 0,5, namas – 0,5, buvo – 0,5, gražus – 0,9, ir – 0,5, jaukus – 0,75, bet – 0,5, mažas – 0.3

- Skaičiuojamos pozityvios ir negatyvios tikimybės

> **Rezultatas:**
> P(T) ≈ 0,7
> P(N) ≈ 0,7
> Kadangi P(T) = P(N), tad pgl. Bing Liu metodiką komentaras nei teigiamas, nei neigiamas.

3.6.2. Patobulintos metodikos žingsniai

- Sumažinamos raidės

> **Rezultatas:** tad tas namas buvo gražus ir jaukus, bet mažas :(

- Normalizuojami jausmukai

> **Rezultatas:** tad tas namas buvo gražus ir jaukus, bet mažas ##0,2##

- Pašalinami skyrybos ženklai

> **Rezultatas:** tad tas namas buvo gražus ir jaukus bet mažas ##0,2##

- Pašalinami pertekliniai žodžiai

> **Rezultatas:** gražus jaukus mažas ##0,2##

- Gaunami kamienai

> **Rezultatas:** graž jauk maž ##0,2##

- Iš mokymo bazės gaunamos tikimybės būti pozityviems

> **Rezultatas:** graž – 0,9, jauk – 0,75, maž – 0.3 ##0,2##

- Skaičiuojamos pozityvios ir negatyvios tikimybės

> **Rezultatas:**
> P(T) ≈ 0,73
> P(N) ≈ 0,9
> Kadangi P(N) > P(N), tad pgl. patobulintą metodiką komentaras neigiamas.

Teoriniame pagrindime pademonstruota, kad nors prof. Bing Liu metodika vykdo mažiau žingsnių, tačiau ji žymi nereikalingus (neutralius) žodžius, ji nepaiso jausmukų ir naudoja

perteklinių duomenų skaičių. Dėl to, kad prof. Bing Liu metodika neatsižvelgia į jausmukus, galima teigti, kad patobulinta metodika yra tikslesnė.

3.6.3. Bing Liu ir patobulintos metodikos mokymo aibių apimties palyginimas

Prof. Bing Liu metodika naudoja didesnę mokymo aibę. Tad tai reiškia, kad apmokyti Bing Liu klasifikatorių reikia gerokai daugiau daugiau. 3 lentelė įrodo, kad žodžiai yra kamieninami, klasifikatoriau reikia nuo kelių iki keliolikos kartų mažiau mokymo duomenų.

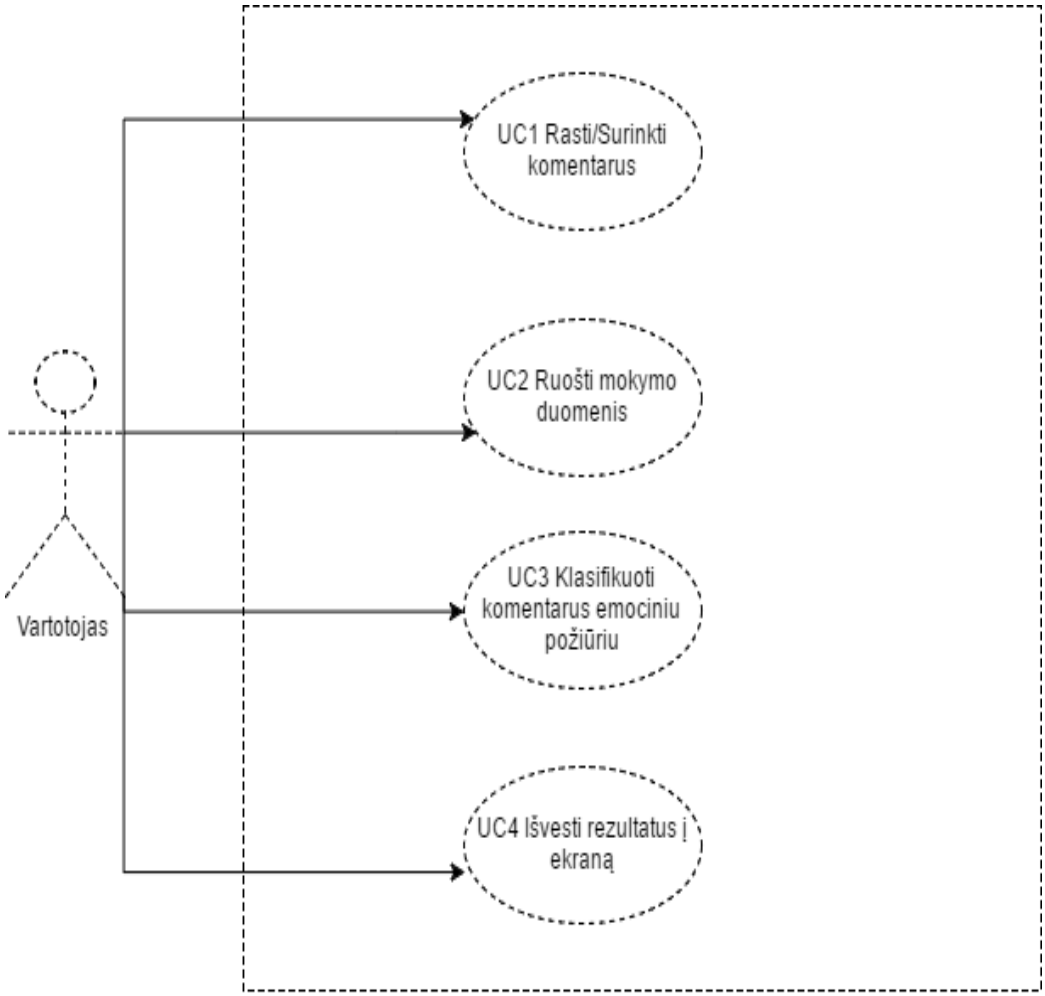
3 lentelė. Mokymo duomenų apimties palyginimas

Bing Liu mokymo duomenys	**Emocinė tikimybė**	**Patobulinta metodikos mokymo duomenys**	**Emocinė tikimybė**
geras	0,9(T)	ger	0,9(T)
gero	0,9(T)		
gerų	0,9(T)		
geri	0,9(T)		
geriausių	0,9(T)		
gėris	0,9(T)		
gerą	0,9(T)		

4. METODIKOS EMPIRINIS PAGRINDIMAS PROTOTIPU

4.1. Prototipo panaudos atvejų diagrama

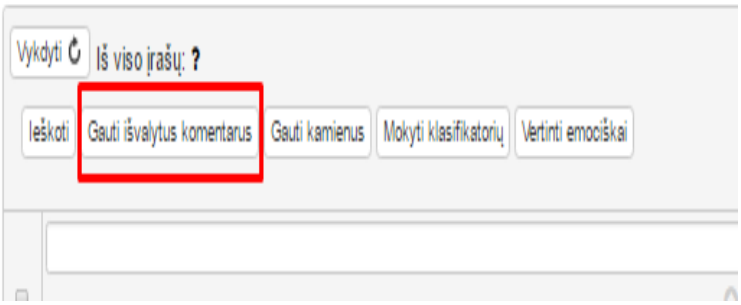
Šiame poskyryje pateikiama prototipo panaudos atvejų diagrama norint parodyti akivaizdžią sąsają tarp patobulintos metodikos ir prototipo (8 pav.).



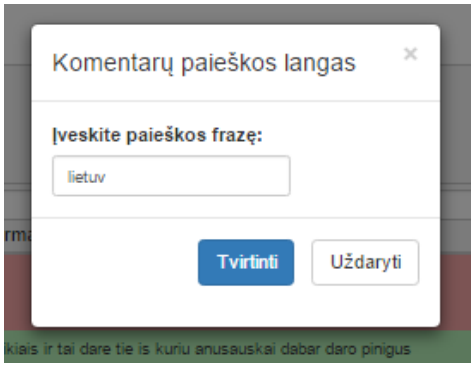
8 pav. Prototipo panaudos atvejų diagrama

4.2. UC1 Rasti/Surinkti komentarus. UC2 Ruošti (mokymo) duomenis

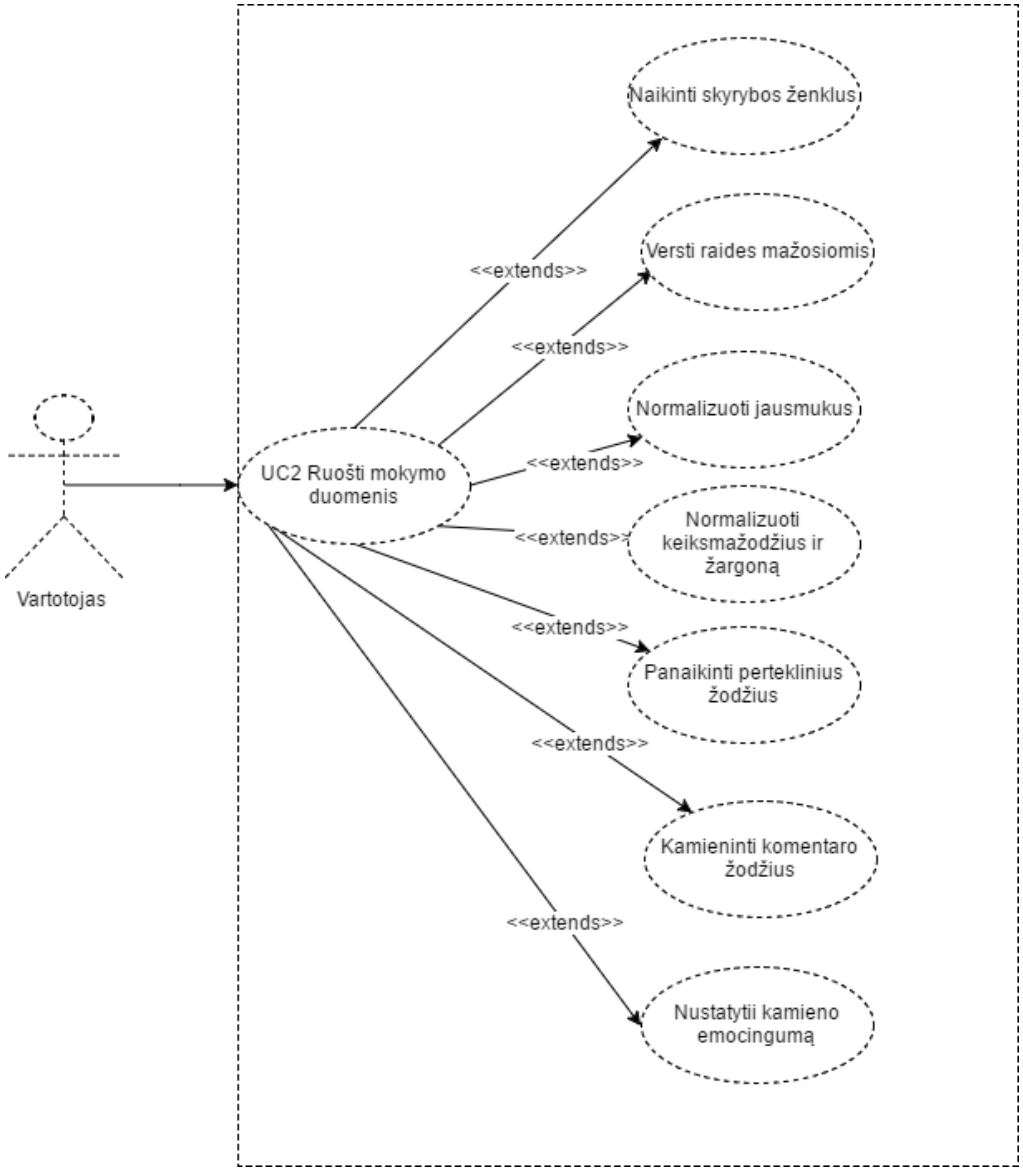
Prototipe šie panaudos atvejai realizuoti surenkant komentarus ir vartotojui pateikiant juos paruoštus (9 pav., 10 pav.).



9 pav. UC1 Rasti/Surinkti komentarus

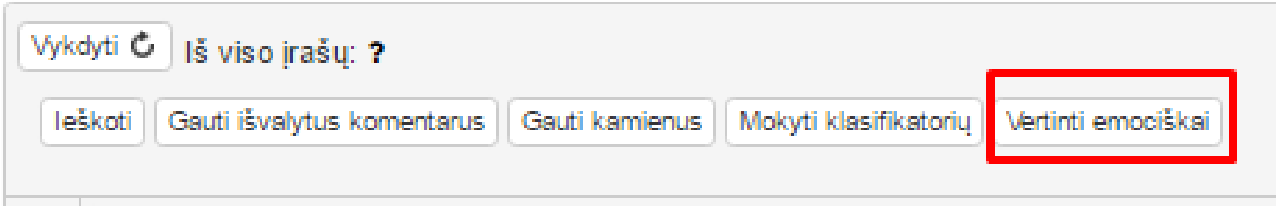


10 pav. UC1 Rasti/Surinkti komentarus



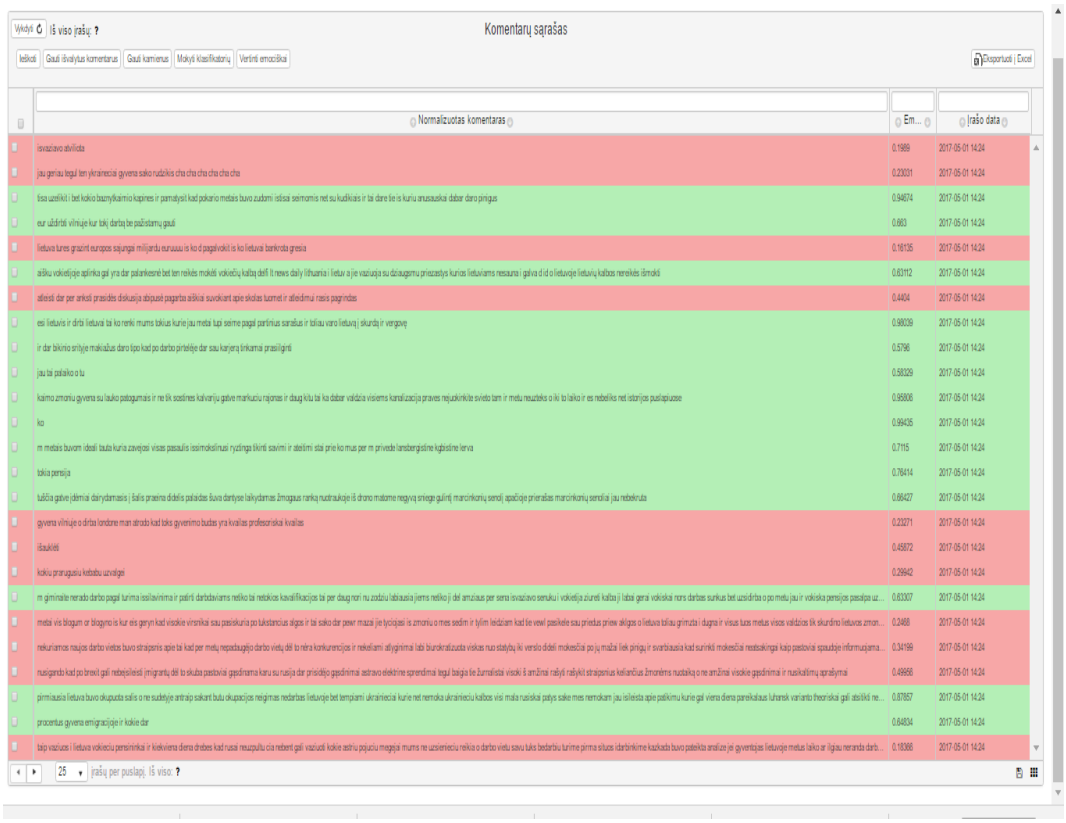
11 pav. UC2 Ruošti mokymo duomenis

4.3. UC3 Klasifikuoti komentarus emociniu požiūriu



12 pav. UC3 Klasifikuoti komentarus emociniu požiūriu

Baigus skaičiuoti komentaro tikimybes priklausyti teigiamai ar neigiamai klasei, komentarai pažymimi teigiamu arba neigiamu atributu, kuris diktuoja, ar komentaras bus nuspalvintas raudonai (neigiamas komentaras), ar komentaras bus nuspalvintas žaliai (teigiamas komentaras) (12 pav.).



13 pav. UC4 Išvesti rezultatus į ekraną

4.4. UC4 Išvesti rezultatus į ekraną

Rezultatai į ekraną išvedami dvejopai: duomenis galima matyti sąraše ir taip juos galima eksportuoti į ekselio failą (13 pav.).

4.5. Prototipo architektūra

Prototipo kūrimui naudotas n-sluoksnių (*angl.* n-tier) programų sistemos architektūros šablonas. Darbo autorius pasirinko kurti aplikaciją, kurią sudaro šie keturi sluoksniai: Web, Services, Domain, Dal (4 lentelė).

4 lentelė. Prototipo architektūros sluoksniai

Prototipo architektūros sluoknis	**Paaiškinimas**
Web	Šiame sluoksnyje yra realizuota kliento dalies logika. Ši logika yra realizuota MVC programų sistemų architektūros šablonu, kurį sudaro MODEL-VIEW-CONTROLLER, t.y. duomenų struktūra, vaizdas ir kontroleris, kuris paruošia duomenis vaizdui.
Services	Šiame sluoksnyje realizuota serverio dalies logika.
Domain	Šiame sluoksnyje deklaruotos prototipo duomenų stuktūrų klasės, kurios skirtos sąsajai su duomenų bazės duomenų struktūromis.
Dal	Šiame sluoksnyje realizuota logika, skirta paimti duomenis iš duomenų bazės ir juos perduoti iš aplikacijos į duomenų bazės.

4.6. Prototipo kūrimui taikytos technologijos

Pirmiausia, prototipo mokymo ir aktualiems analizės duomenims saugoti naudota struktūriška MS SQL reliacinė-objektinė duomenų bazė.

Antra, prototipo struktūra sukurta su Asp.Net MVC karkasu. Prototipo serverio dalies logika kurta su C# kalba ir jos Linq to SQL biblioteka, kuri leidžia naudoti Object Relational Mapping metodą, įgalinantį linq kalbos konvertaciją į SQL užklausas. Prototipo kliento dalis pateikiama HTML DOM‘u, kurį sudaro HTML, CSS, Javascript žymių, taisyklių ir programavimo kalbos. Prototipą taip pat papildo tokios javascript’o bibliotekos - Jquery, AJAX, RequireJs, jsZip [5 lentelė].

5 *lentelė. Prototipo javascript bibliotekų naudojimo paaiškinimas*

Javascript‘o bibliotekos	**Naudojimo motyvacija**
jQuery	Ši biblioteka padeda lengviau manipuliuoti html objektais.
Ajax	Ši biblioteka padeda komunikuoti su serveriu be e. svetainės atnaujinimo.
RequireJs	Šia biblioteka script‘ai moduliarizuojami ir taip išvengiama tokių blogų praktikų, pavyzdžiui, globalūs kintamieji.
jsZip	Šia biblioteka Kendo grid‘o sąrašai konvertuojami į Microsoft excel‘io ataskaitą.

4.7. Eksperimentas metodikos tikslumui įvertinti

Eskperimento metodikos tikslumo vertinimas prototipu vykdytas taip: buvo paimti atsitiktiniai surinkti ir prototipo suklasifikuoti komentarai [6 lentelė]. Tada buvo lyginamas klasifikatoriaus vertinimas su eskpertinėmis išvadomis, ar komentaras labiau pozityvus ar negatyvus. Didžiąja dalimi – apie 67,5 - komentarų emocijų vertinimas buvo tikslus.

Toks tikslumo skaičius yra aukštas, nes, pirma, buvo klasifikuojami komentarai, kurie ilgis būna nuo sakinio lygio iki dokumento lygio (10 sakinių). Jei komentarų ilgis būna 1-2 sakiniai, tuomet klasifikavimo tikslumas gali siekti 80 proc. [RG2003], o gal net ir viršyti. Darbo autorius pritaria straipsnio [RG2003] prielaidoms, kuriomis teigiama, kad 100 proc. tikimybės pasiekti neįmanoma, nes mokymų duomenis kuria žmonės (angl. „relies on human categorization“), o tai jau neigiamas žmogiškasis faktorius.

6 lentelė. Eksperimento duomenys

Matavimo numeris	**Matavimo tiksli dalis**	
Matavimas Nr. 1	15/20	75 %
Matavimas Nr. 2	16/20	80 %
Matavimas Nr. 3	14/20	70 %
Matavimas Nr. 4	11/20	55 %
Matavimas Nr. 5	13/20	65 %
Matavimas Nr. 6	16/20	80 %
Matavimas Nr. 7	11/20	55 %
Matavimas Nr. 8	16/20	80 %
Matavimas Nr. 9	12/20	60 %
Matavimas Nr. 10	13/20	55 %

REZULTATAI IR IŠVADOS

Darbo metu buvo gauti tokie rezultatai:

1. išanalizuoti emocijų aptikimo tekste metodai nustatant metodus, tinkamus anoniminių e. portalų komentarų emocijų tyrybai;
2. sukurtas patobulintos e. portalų komentarų tyrybos metodikos prototipas, kurio tikslumas, atlikus ekperimentą, siekia 67,5 proc.;
3. sukurta klasifikatoriaus mokymo bazė;
4. patobulinta prof. Bing Liu emocijų tyrybos metodika:
    - pagal šiuolaikinę komentarų analizės praktiką pridėtas mokymo duomenų paruošimo komponentas su duomenų paruošimo metodais;
    - pagal šiuolaikinę komentarų analizės praktiką į klasifikavimo komponentą įeina ir mašininio mokymo metodas (Bayes Multinomial Classifier), ir leksikono metodas.

Darbo metu buvo gautos tokios išvados:

1. atlikus eksperimentą, gautas tikslumas, kuris yra pakankamas;
2. teoriškai įrodyta, kad įvestas kamieninimo metodas leido sumažinti mokymo duomenų apimti nuo kelių iki keliolikos kartų;
3. metodika gali būti taikoma įvairių temų komentarų įvertinimui emociniu požiūriu sudarius klasifikatoriaus mokymo duomenų bazę, paremtą ekspertinėmis išvadomis.

ŠALTINIAI

1. [Agr2012] Agrawal Ammeta, *Unsupervised Emotion Detection from Text using Semantic and Syntactic Relations*. In: WI-IAT '12: Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology, Vol. 1, p. 346-353. E. prieiga: www.cs.yorku.ca/~aan/research/paper/Emo_WI10.pdf
2. [ARS2005] C. O. Alm, D. Roth, and R. Sproat *Emotions from text: machine learning for text-based emotion prediction*. In: Proceedings of HLT/EMNLP, 2005, p 579–586.
3. [Bin2010] Binali Haji *Computational Approaches for Emotion Detection in Text*. In: 4th IEEE International Conference on Digital Ecosystems and Technologies, 2010, 172-177.
4. [Can2014] Canales Lea *Emotion Detection from text: A Survey*, 2014. E. prieiga: http://repository.dlsi.ua.es/936/1/Emotion_Detection_from_text_A_Survey_revisado.pdf.
5. [Car2006] Carletta J. et al. The AMI *Meeting Corpus: A Pre-announcement*. In: Renals S., Bengio S. (eds) Machine Learning for Multimodal Interaction. Lecture Notes in Computer Science, vol 3869. Springer, Berlin, 2006.
6. [Cha2012] Chalothorn Tawunrat *Using SentiWordNet and Sentiment Analysis for Detecting Radical Content of Web Forums*. In: 6th Conference on Software, Knowledge, Information Management and Applications (SKIMA 2012), 2012, p. 9-11.
7. [CK2013] Calvo R.A., Kim M. *Emotions in text: dimensional and categorical models*, Vol. 29(3), p. 527–543, 2013.
8. [Des2013] Deshpande Shriniwas Approaches towards Emotion Extraction from Text. In: National Conference on Innovative Paradigms in Engineering & Technology Proceedings, 2013, p. 10-14.
9. [Dob2001] Dobrescu Alexandra Balahur ,,Methods and Resources for Sentiment Analysis in Multilingual documents of Different Text Types“, University of Alicante, 2001: Daktaro disertacija.
10. [Eck1999] Ekman, Paul *Basic Emotions*. In Dalgleish, T. & Power, M. J. (Eds.), Handbook of Cognition and Emotion (pp. 45-60). New York, 1999.
11. [Gil2005] Gill Alastair J. *Texttone - Emotion detection in short texts*, 2005. E. prieiga: homepages.inf.ed.ac.uk/jon/papers/texttone1.pdf

12. [Yam2015] Yam Chew-Yean *Emotion Detection and Recognition from Text*, 2015. E. prieiga: https://www.microsoft.com/developerblog/real-life-code/2015/11/30/Emotion-Detection-and-Recognition-from-Text-using-Deep-Learning.html
13. [Jur2014] Jurafsky Dan *Sentiment Analysis. What is Sentiment Analysis?*: Skaidrės. E. prieiga: https://web.stanford.edu/class/cs124/lec/sentiment.pptx
14. [Kim2011] Kim Sunghwan Mac *Recognising Emotions and Sentiments in Text*, University of Sydney, 2011: Magistro darbas.
15. [Liu2011] Liu Bing *Sentiment Analysis and Opinion Mining*, Illinois, 2011.
16. [Med2014] Medhat Walaa *Sentiment analysis algorithms and applications: A survey*. In: Ain Shams Engineering Journal, Vol. 5, 2014, 1093–1113.
17. [Mic2013] Michael Christina *Sentiment Analysis For Debates*, Imperial College London, 2013: Daktaro Disertacija
18. [Mor2012] Morris Travis *Extracting and Networking Emotions in Extremist Propaganda*. In: 2012 European Intelligence and Security Informatics Conference, 2012, p. 53-59.
19. [Mul2013] Mullen R. Introduction to Sentiment Analysis, Saarland University: skaidrės. E. prieiga: http://lct master.org/files/MullenSentimentCourseSlides.pdf
20. [RG2003] Rambocas Meena , Gama João *Marketing Research: The Role of Sentiment Analysis*. In: FEP WORKING PAPERS, 2003.
21. [RMN2012] RC Balabantaray, Mudasir Mohammad, and Nibha Sharma *Multi-Class Twitter Emotion Classification: A New Approach*. International Journal of Applied Information Systems (IJAIS), p. 48–53, 2012.
22. [RP2013] Ringsquandl Martin, Petković Dušan *Analyzing Political Sentiment on Twitter*. In: Analyzing Microtext: Papers from the 2013 AAAI Spring Symposium, p. 40-47.
23. [RRJ2012] Roberts Kirk, Roach Michael A., Johnson Joseph, Guthrie Josh, Harabagiu Sanda M. *Empatweet: Annotating and detecting emotions on twitter*. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012). P.30-39, 2012.
24. [SI2013] Suttles J., Ide N. Distant Supervision for Emotion Classification with Discrete Binary Values. In: Gelbukh A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2013. Lecture Notes in Computer Science, vol 7817. Springer, Berlin, Heidelberg

25. [SM2007] C. Strapparava and R. Mihalcea *SemEval-2007 task14: Affective Text*. In:Proceedings of SemEval-2007, Prague, Czech Republic, 2007.
26. [SRW2007] Somasundaran Swapna, Ruppenhofer Josef, Wiebe Janyce *Detecting Arguing and Sentiment in Meetings*, SIGdial Workshop on Discourse and Dialogue, Antwerp, Belgium, September 2007.
27. [Web2017] Webster'io anglų kalbos žodynas. E. prieiga: http://www.merriam-webster.com/dictionary

SĄVOKŲ APIBRĖŽIMAI

1. E. portalai – naujienų portalai.
2. E. portalų komentarai – nuomonės po naujienų portalų straipsniais.
3. Komentarų tyryba/sentimentų analizė – duomenų gavybos disciplina, kurios tikslas gauti subjektyvią informaciją iš interneto tekstų, ypač interneto komentarų.
4. Metodika - būdų, taisyklių visuma kuriam nors darbui gerai atlikti.
5. Prof. Bing Liu – Ilinojaus universiteto profesorius, vienas iš labiausiai cituojamų sentimentų analizės autoritetų pasaulyje.