TARTU ÜLIKOOL
MATEMAATIKA-INFORMAATIKATEADUSKOND
Arvutiteaduse instituut
Infotehnoloogia eriala

Siim-Toomas Marran

Sentimentaalne analüüs eestikeelse peavoolumeedia veebiartiklite kommentaaride baasil

Bakalaureusetöö (6 EAP)

Juhendaja: Peep Küngas
Kaasjuhendaja: Meelis Kull

Autor: ……..… "….." mai 2012
Juhendaja: ……..…. "….." mai 2012
Juhendaja: ……..…. "….." mai 2012

Lubada kaitsmisele
Professor ..... ……. "….." mai 2012

TARTU 2012

Sisukord

1 Sissejuhatus

Iga inimene saab levitada informatsiooni veebis, näiteks avaldada arvamust erinevatel poliitilistel teemadel, väljendada rahulolematust erinevate toodete üle, juhtida tähelepanu teenindussektoris olevate puudujääkide poole ja nii edasi ning kanaleid, mille kaudu teavet edasi anda isegi rohkem, näiteks foorumid, blogid, mikroblogid, uudiste kommentaariumid, sotsiaalportaalid, wikid, massiivsete kasutuskondadega veebimängud jne. Viimane väide on pigem metafooriline, aga sellegipoolest sotsiaalmeedia on viimastel aastatel arenenud hüppeliselt, pidevalt luuakse uusi rakendusi, kus osade teenuste kasutajate hulk küündib sadadesse miljonitesse, näiteks suhtlusportaal Facebook omab aprill 2012 aasta seisuga 901 miljonit kasutajat [1], mikroblogimiskeskond Twitter omab 2012 aasta märtsi seisuga 140 miljonit kasutajat [2]. Selline kasutajate hulk tähendab tohutut andmehulka, päevas toodetakse miljardeid postitusi ning paljud postitused omavad paljudele teistele isikutele, firmadele, organisatsioonidele ja valitsustele tohutut väärtust.

Tohutu huvi informatsioonitulva vastu tunnevad nii teadusringkonnad kui eraettevõtted, kes tegelevad andmekaeve ning nende andmete analüüsiga. Põhjus, miks on huvi tärganud seisneb selles, et informatsioon omab majanduslikku väärtust. Selle bakalaureusetöö raames keskendume automatiseeritud andmekaevele ning kogutud informatsiooni sentimentaalsuse hindamisele. Sentimentaalsuse hindamine tähendab andmete polaarsuse ehk lihtsamalt öeldes positiivsuse ja negatiivsuse määramist. Näiteks hinnatakse fraas „*tark* laps“ positiivseks ning „*vihane* tudeng“ negatiivseks, nendes fraasides polaarsust mõjutavad sõnad on kaldkirjas.

Töö eesmärk on kombata peamisi Eesti uudisteportaale, mis on veebikasutajate seas levinud: Postimees, Äripäev, Ekspress, Päevaleht ja Delfi. Rakendus tegeleb eelmainitud lehekülgedel andmekaevega ning salvestab andmebaasi andmed, milleks on artiklite kommentaarid. Hiljem on võimalik kasutajatel loodud veebilehel ennast huvitavatel teemadel korraldada päringuid. Päringu ajal kogutakse andmebaasi salvestatud informatsioon ja kuvatakse töödelduna kasutaja ekraanile sektordiagrammina. Sektordiagrammi abil väljendatakse otsingutulemuse kohta käivate kommentaaride olemust protsentuaalselt. Kommentaaride olemust saab näidata kolmel viisil: positiivne, negatiivne ja neutraalne. Kogu seda kogumise ja hindamise protsessi nimetatakse sentimentaalseks analüüsiks.

2 Sentimentaalne analüüs

2.1 Definitsioon

Sentimentaalne analüüs on ülesanne identifitseerida positiivseid ja negatiivseid arvamusi, emotsioone ning hinnanguid [3]. Sentimentaalse analüüsi hinnang ei pruugi alati olla eelnevalt nimetatud, kas heakskiitev või laitev, vaid võib olla ka mõlemat või hoopiski neutraalne. Identifitseeritavate objektide sentimentaalset väärtust nimetatakse polaarsuseks. Näiteks allpool on toodud fraasid, kus kaldkirjas on näidatud fraasile polaarsuse andvad sõnad või alamfraasid:

(1) „*Tubli* inimene“ – positiivne fraas;
(2) „Totaalne *katastroof*“ – negatiivne fraas;
(3) „*Keskpärane* esitus“ – neutraalne fraas.

Viimast loeme neutraalseks, sest olenevalt kontekstist ja subjektiivsest hinnangust võib fraas „*keskpärane* esitus“ omada nii positiivset kui negatiivset väärtust. Sentimentaalset analüüsi on võimalik läbi viia nii inimeste poolt(hindajad on inimesed) kui ka automaatselt ehk arvutusmeetodeid kasutades. Viimase puhul on tegemist kiirelt areneva teadusharuga, sest vajadus automatiseeritud hindamisele on suur ning põhjendatud(vt. alapeatükk 2.3). Inimeste poolt antud hinnangud pole objektiivsed, sest avaldavad inimeste arvamust ja tõekspidamisi seepärast on tegu subjektiivse hindamisega. Samas ei saa väita, et automatiseeritud analüüs tagastaks objektiivseid hinnanguid, kuna rakendustes kasutatavad leksikonid(vt. alapeatükk 2.2) on enamasti koostatud inimeste poolt. Seetõttu kutsutakse seda ka subjektiivseks analüüsiks.

Sentimentaalse analüüsi poole võib pöörduda ka mõistega *„opinion mining“* ehk otsetõlkes arvamuse kaevamine, mis kujutab endas automatiseeritud arvamuste kogumist ja hindamist. Tegemist on rakendusega, mis koosneb erinevatest meetoditest, et eraldada subjektiivset teavet lähtematerjalist [4].

Semantiline orientantsioon näitab sõna hinnatava karakteri kuuluvust. Semantiline orientatsioon määrab sõnade polaarsust, kas neutraalne, negatiivne või positiivne. Sellisel juhul positiivsed sõnad näitavad ihaldusväärseid omadusi(„ausus“, „kartmatu“) ning negatiivsed sõnad mitte-ihaldusväärseid(„häiriv“, „vihkav“) [5]. Näiteks, kuidas semantiline orientatsioon praktikas töötab, võetakse keeleline korpus, kust valitakse kõik omadussõnad, mis esinevad sagedusega 20 või enam korda korpuses. Eemaldatakse neutraalsed omadussõnad ning ülejäänud klassifitseeritakse, kas positiivseteks või negatiivseteks. Semantiline orientatsioon tugineb paljuski masinõppes esinevate meetoditele, et kuidas sõnade tähenduslikku väärtust hinnatakse. Käesolevas töös meie hindame sõnade polaarsust käsitsi ning määrame kahte erinevasse „hunnikusse“(positiivsed ja negatiivsed) ehk semantiline orientatsioon „kott sõnadega“ meetodil.

2.2 Ressurssid

Sentimentaalse analüüsi puhul on vaja mitmeid ressursse. Peamine ülesanne on koguda sotsiaalmeediast brändide ja isikute kohta käivat informatsiooni. Teabe allikateks võib kasutada erinevaid suhtlusportaale nagu Facebook, Twitter, Google+, uudiste portaale, kommentaariume, foorumeid, blogisi, arvustussaite ehk veebilehekülgi ja dokumente, kus inimesed jagavad subjektiivset informatsiooni, mille sentimentaalsust saab hinnata. Andmete kogumine on üldjuhul automatiseeritud, enamasti kasutatakse selleks veebiroomajaid(vt. peatükk 3).

Kui hindamist vajav andmestik on loodud, tekib küsimus, et mismoodi on üldse võimalik määrata mõne fraasi või sõna polaarsust? Selle jaoks on vaja luua polaarsusleksikone, mis on üldjuhul andmebaasid, mis omavad positiivseid ja negatiivseid sõnu ja fraase, mille abil on hiljem võimalik määrata konteksti polaarsust. Mõiste polaarsusleksikon pole üldtuntud mõiste, vaid käesolevas töös autori poolt kasutusele võetud mõiste leksikonide ehk sõnakogumite jaoks, mis koosnevad polaarsust näitavatest sõnadest ja fraasidest. Näiteks polaarsusleksikon omab positiivseid sõnu („tark“, „tubli“, „ilus“) ja negatiivseid sõnu („mage“, „inetu“, „jube“). Hindamiseks antud kontekst: „Mallel on jube harjumus inimesi sõimata.“ Olenemata hindamisalgoritmist ja –meetodist on teada, et kontekst omab ühte negatiivset sõna.

Viimane tähtis ressurss on algoritm või meetod, et kuidas lahendada probleem ehk mismoodi hinnata veebist kaevandatud subjektiivse teavet.

2.3 Vajadus

Põhjus, miks sentimentaalne analüüs, teise nimega arvamuskaeve, on nii populaarseks muutunud, seisneb selles, et inimesed on hakkanud rohkem avaldama Internetis oma arvamust, andma hinnanguid ja väljendama emotsioone. Informatsioon, mida inimesed levitavad, on hakanud huvi pakkuma teistele inimestele, firmadele ja valitsustele.

Inimesed otsivad Internetist toote tutvustusi, et kontrollida toote kvaliteeti. 81% veebi kasutajatest on uurinud produkte veebist vähemalt korra elus ning 20% teevad seda igapäevaselt [6].

Poliitikute huvi selle areneva teadusharu vastu seisneb selles, et koguda ühiskonna tagasisidet poliitikute suhtes – näiteks imidži uuring. Valitsuse salaluurel on kasulik leida kindlate otsingusõnade ja viidete abil võimalikke ohte ja negatiivseid infovahetusi.

Firmad on avastanud, et Internet on ideaalseks kohaks, et kust saada tagasisidet oma toodete kohta ilma et peaks läbi viima kalleid turuuuringuid. Ettevõte saab näha, miks keegi ei soovi soetada nende poolt toodetud sülearvutit ning saavad tagasisidet, kuidas nende toodet parandada ja muuta konkurentsivõimelisemaks. Lisaks, Internetis ülevalolevad toodete ja majutus kohtade arvustused mõjutavad oluliselt ettevõtete

käekäiku. Kliendid on valmis toodete eest 20-99% rohkem maksma, millel 5 tärni retsentsioonis, kui 4 tärniste. 73-87% turistidest tunnistavad, et arvustused mõjutavad nende oste [6].

Sentimentaalne analüüs aitab luua tingimusi, et kiirelt ning edukalt koguda teavet veebist ja töödelda. Enam pole alati tarvis läbi viia kalleid turuuuringuid, mille käigus küsitletakse tuhandeid inimesi ning mille tulemusi tuleb sisestada andmebaasi ning statistliselt töödelda. Uued lahendused on üldjuhul automatiseeritud, seetõttu pikemas perspektiivis ka odavamad. Peamine motivatsioon on siiski majanduslik. Nimelt peitub tänapäeval informatsioonis võim ja jõud ning informatsioon omab kõrget väärtust – tegu on infoajastuga, kus majandusliku eelise annab teadmine.

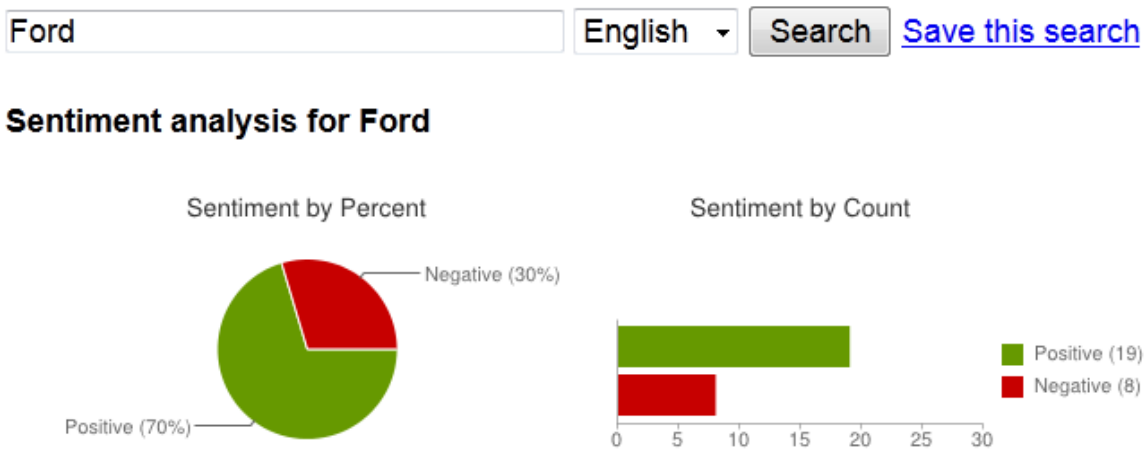
2.4 Olemasolevad tööriistad

Praegusel hetkel turul olevad tooted pakuvad reaalajas sotsiaalmeedia jälgimist, brändi-imago ülevaateid, rakendusi, mida integreerida oma kodulehe või intraneti ehk sisevõrguga, ja palju muid analüütilisi funktsioone ning uuringuid. Enamasti erinevad tooted keskenduvad massimeediale nagu Twitter ja Facebook, mis omavad miljoneid kasutajaid ehk miljardeid postitusi päevas. Kuulsamad ja populaarsemad rakendused ja teenuse pakkujad:

(1) Radian6 [7], tuntumad kliendid: Air Canada, Pepsi, L'Oreal, UPS, 3M, Activision, Logitech;
(2) Lithium [8], tuntumad kliendid: TomTom, AT&T, Lenovo;
(3) MediaVantage [9], tuntumad kliendid: BurgerKing, McDonalds, Royal Bank of Canada.

Head näited, millised näevad välja sentimentaalse analüüsi rakendused ning mismoodi tulemusi kuvatakse, on kõigile vabalt kättesaadavad reaalajalise sotsiaalmeedia otsingu- ja analüüsi rakendused:

(1) Socialmention* [10],
(2) Sentiment140 [11].



Joonis 1:näide rakenduse Sentiment140 kohta, päringuks „Ford“ bränd.

Joonisel 1 on näha, Sentiment140 kohta tehtud näide. Näites on otsitud informatsiooni Ameerika Ühendriikide autotootja „Ford“ kohta. Joonisel on näha, et kahte erinevat tüüpi diagrammi: sektor- ja tulpdiagramm. Diagrammidel kuvatakse positiivsete ja negatiivsete Twitter-postituste omavahelist suhet. Sektordiagrammil protsentuaalselt ning tulpdiagrammil on kujutatud arvuliselt, et kui palju positiivseid ja negatiivseid postitusi esines.

Käesoleva bakalaureusetöö implementatsioon on võrreldes vabavaraliste ning avatud lähtekoodiga teostega võrdväärne. Võrdväärsus seisneb selles, et tegu on automaatiseeritud andmekaeve- ja hindamisrakendusega, mis ei oma erilisi lisafunktsioone. SocialMention* puhul on tegemist väga laiahaardelise projektiga, mis kogub informatsiooni väga paljudest erinevatest sotsiaalmeedia kanalitest, samal ajal Sentiment140 piirdub ainult Twitteriga. Käesolev rakendus on spetsialiseerunud eestikeelsele sotsiaalmeediale – uudisteportaalidele, mis on iseenesest eelmainitud rakendustega võrreldes spetsialiseeruv lähenemine.

2.5 Peamised probleemid ja negatiivsed küljed

Majanduslikust vaatenurgast vaadatuna seisneb peamine probleem selles, et inimesed levitavad mõnikord sihilikult valeinformatsiooni sotsiaalmeedia kanalites. Sentimentaalse analüüsi rakendused pole võimelised hindama andmete allika tõepärasust. Statistiliselt on võimalik keskväärtuse suhtes liiga suure hälbega olevad andmed eemaldada, kuid iga lähenemise ja meetodi puhul pole see võimalik.

Sentimentaalse analüüsi rakendused pole võimelised arusaama slängist, ega mõistma sarkasmi. Lisaks erinevad polaarsusleksikone loovate inimeste subjektiivsed arvamused

olenevalt nende silmaringist. Mis on mõne jaoks positiivse tähendusega on teisele negatiivne, ja vastupidi.

3 Veebiroomaja

3.1 Definitsioon

Veebiroomaja on arvutiprogramm või programmide kogum, mis lehitseb meetodilisel, automaatsel viisil või teatud nõutete kohaselt *World Wide Web*'i. Veebiroomajad laevad alla veebilehti, koguvad URL'e HTML keskkondadest. URL'i puhul on tegemist stringiga, mis viitab mingisugusele veebi ressurssile. Roomajaid võib kutsuda ka terminitega sipelgad, automaatsed indekseerijad, *bot*'id, veebiämblikud, veebirobotid.

Veebi arenguga on kasvanud akadeemiliste uurijate ja asutuste huvi veebi vastu. Internet peidab endas suurel hulgal informatsiooni ning selle töötlemiseks on vaja suuremõõtmelist töötlust. Veebi lehekülgede ettesöötmiseks analüüsirakendustele on vaja veebiroomajat, mis indekseeriks ehk koguks olemasolevaid lehekülgi. [12]

3.2 Kasutusvaldkonnad

3.2.1 Otsingumootorid

Üks kõige tähtsamaid rolle veebiroomajatel on toetada otsingumootorite tööd, kus roomajad liiguvad lehekülgedelt lehekülgedele, samal ajal abistavad programmid indekseerivad veebilehti.

Näiteks *Googlebot*, mille puhul on tegemist ülimalt võimeka veebiroomajaga, sest erinevalt paljudest teistest roomajatest on see võimeline jooksutama Javascript'i ning on võimeline töötlema AJAX'i päringuid. Sedasi on veebiroomajad võimelised indekseerima dünaamilist Interneti keskkonda.

3.2.2 Andmekaeve

Veebi andmekaeve ehk *web data mining* puhul on tegemist tehnikaga, mille puhul roomatakse läbi erinevaid veebiressursse ning kogutakse vajalikku informatsiooni. Kõik see võimaldab firmadel arendada ettevõtlust, mõtestada lahti turu dünaamikat jne. Andmekaeve on üldjuhul üks komponent suuremast rakendusest. Kogutud andmestikku saab töödelda paljudel erinevatel viisidel nagu näiteks matemaatiliselt, statistiliselt, keeleliselt jne.

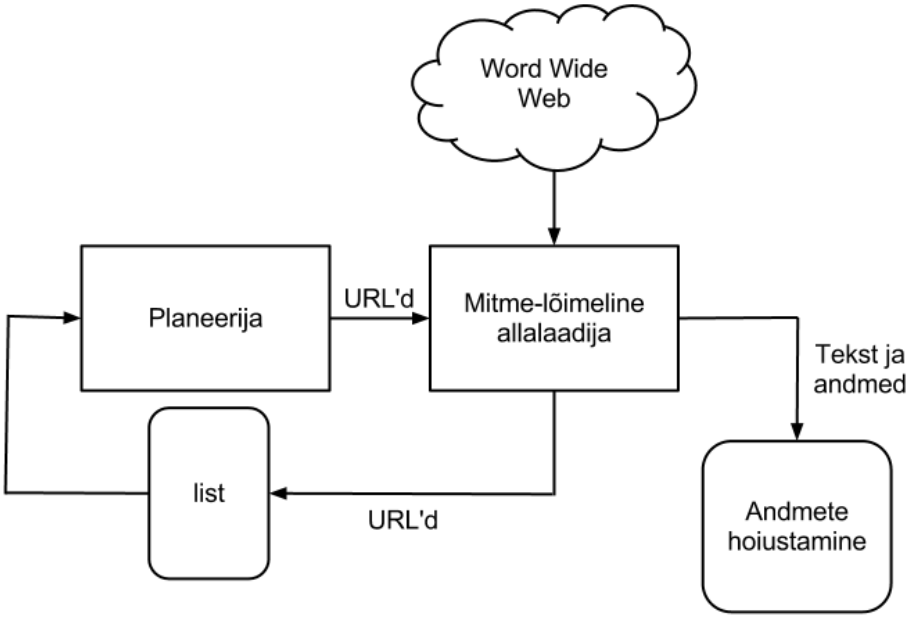
3.3 Tööpõhimõte ja strateegiad

Lihtsustatud veebiroomaja algoritm on suhteliselt triviaalne:

(1) Sisestatakse nimekiri URL'idega, mida soovitakse külastada, programm loob nimekirja alusel listi ehk külastatavate URL'de list ;
(2) Luuakse list, kuhu pannakse URL'd, mida on juba külastatud, esialgu on list tühi;
(3) Alustatakse tsükliga, mis kestab niikaua kuni külastatavate URL'de list ei oma enam elemente;
(4) Külastatavate URL'de listist võetakse element ning eemaldatakse listist;
(5) URL lisatakse külastatud elementide listi;
(6) Elementi ehk URL'i parsitakse ning kogutakse uusi URL'e;
(7) Uued URL'd kontrollitakse läbi, et ega nad ei kuulu olemasolevatesse listidesse(külastatud või külastamata). Kui nad ei kuulu ühtegi listi, siis lisatakse nad külastatavate URL'de listi. [13]

Etteantud nimekirjaga URL'ide sügavuseks määratakse 0 ning igal järgneval sammul saavad uued URL'd ühe võrra suurema sügavuse kui eelmisel etapil.

Tüüpilise veebiroomaja arhitektuur graafilisel kujul:



Joonis 2: Veebiroomaja arhitektuur [13]

Mitme-lõimeline allalaadija laeb veebist alla veebilehti, mis on planeerija talle ette söötnud samal ajal abi programmid eraldavad ning hoiustavad andmeid ning allalaadija lisab listi uusi külastatavaid URL'e.

Veebiroomajate käitumine määratakse strateegiate abil, millest levinumad on järgnevad:

(1) Valiku strateegia ehk mis lehekülgi allalaadida;

(2) taaskülastusstrateegia ehk milliseid lehekülgi taaskülastada, et jälgida muutusi;
(3) viisakusstrateegia ehk kuidas vältida lehtede ülekoormust;
(4) paralleelsusstrateegia ehk kuidas koordineerida hajutatud veebiroomajaid. [13]

Strateegiaid rakendatakse, et vältida võimalikke takistusi ja probleeme.

3.4 Peamised takistused, probleemid ja lahendused

Olenevalt veebiroomaja ülesandest tuleb ühel hetkel päevakorda tõsiasi, et missuguseid lehekülgi tasub läbida ning milliseid mitte. 2005. aastal suuremad otsingumootorid indekseerisid 40-70% kogu veebist [14]. Siin jõuame tõdemusele, et kui isegi suurkorporatsioonid pole võimelised oma võimsa masinapargiga tervet veebi indekseerima, siis väikearendajatel tuleb ilmtingimata teha valikuid, missuguseid lehekülgi lehitseda. Seetõttu kasutatakse roomamisel erinevaid algoritme ning parameetreid, näiteks erinevad sügavuti- ja laiutiotsingu algoritmid.

Otsingumootoritele on ülitähtis, et nende poolt pakutav informatsioon oleks õige ning ajakohane. Sellepärast määratakse veebiroomajatele strateegiatega kohustus, mille kohaselt tuleb uuesti külastada juba roomatud lehekülgi. Sedasi avastatakse lehekülgedel toimunud uuendusi nagu lisatud või eemaldatud URL'd. Otsingumootoritele pole midagi hullemat kui levitada vananenud informatsiooni.

Paljud veebiadministraatorid on täheldanud, et veebiroomajad koormavad liialt võrguliiklust ning seetõttu on võetud kasutusele viisakusstrateegia, mille käigus arendajad saavad anda roomajatele teada, et kas lehekülge lubatakse roomata või mitte.

Veebiroomaja efektiivsust ja kiirust pärsib tugevalt erinevate veebilehtede masinatelt vastuste saamine ehk pikk ooteaeg. Seetõttu kirjutatakse roomajaid mitme-lõimelistena. Selle tulemusena saab üks protsess korraga taotleda informatsiooni sadadelt veebilehekülgedelt.

4 Implementatsioon

Rakenduseks nimetame kogu bakalaureusetöö raames arendatavat tarkvara lahendust, mis koosneb andmekaevest ehk kommentaaride kogumisest sotsiaalmeediast. Kogutud kommentaarid hinnatakse ja tulemus kuvatakse kasutajale. Rakendus kättesaadaval lisades(vt. alapeatükk 10.2).

4.1 Rakenduses kasutatavad tehnoloogiad

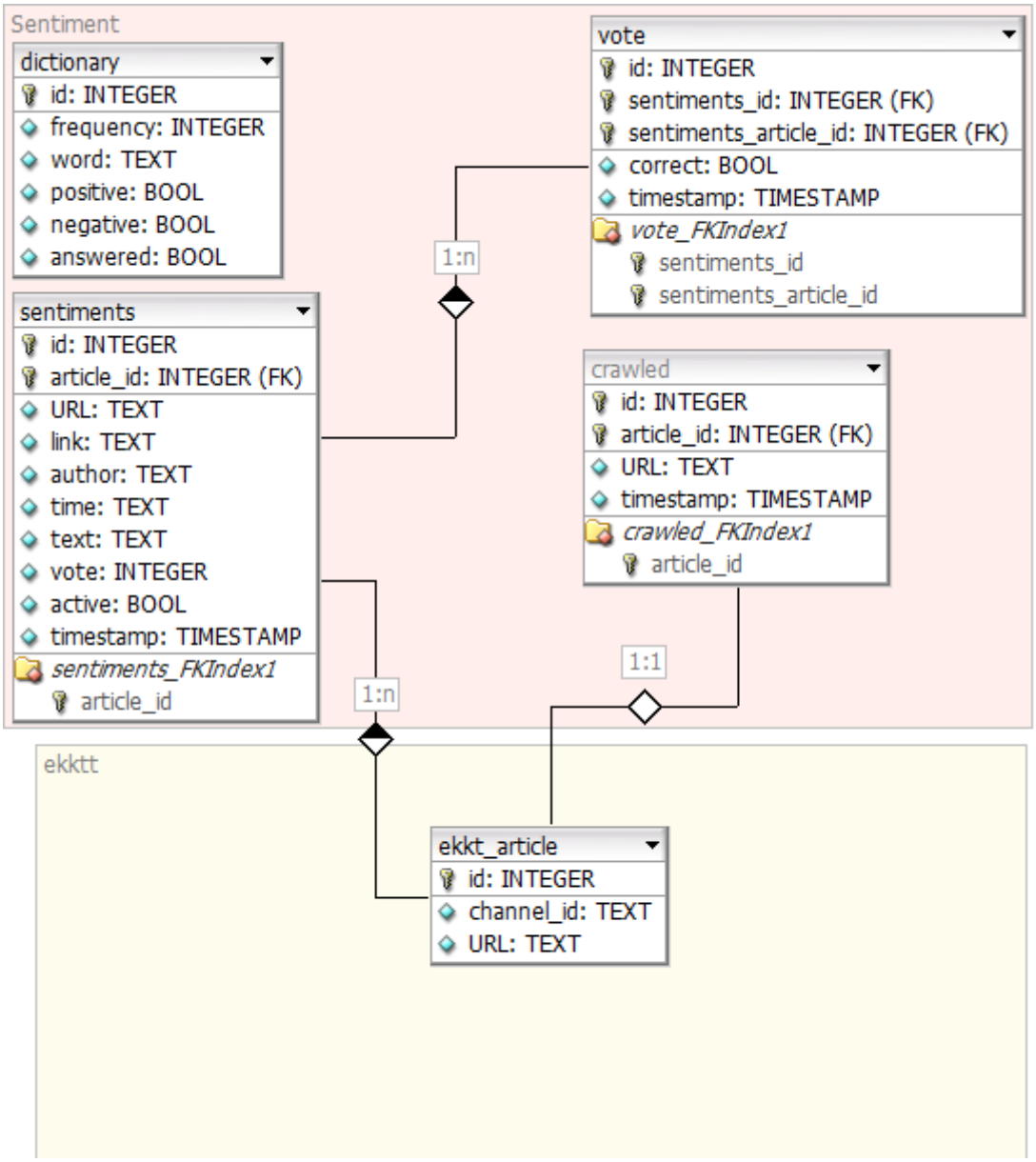
Käesoleva tarkvaraga on tegu veebirakendusega, mis nõuab praktikas teatud infrastruktuuri olemasolu(veebiserveri tarkvara, andmebaas, programmeerimiskeele tugi). Sellepärast võtsin kasutusele programmide koondprodukti WAMP, mis tähendab Windows-Apache-MySQL-PHP. Apache HTTP Serveri puhul on tegu veebiserveri tarkvaraga, mis on ülimalt populaarne üle maailma, hetkel kasutab ligikaudu 57,5% kõikidest serveritest Apache tarkvara [15].Sellise infrastruktuuri lahenduse kasutuselevõtu üheks põhjuseks on ka kindlasti MySQL ja PHP, mis annavad arendajatele lihtsa viisi luua veebirakendusi, sest MySQL on tüüpiline relatsioonilise andmebaasi keele SQL esindaja ning vabavaraline. Samal ajal PHP omab kasutajasõbralikku MySQL teeki. WAMP'i puhul on tegemist kergelt konfigureeritava tootega, lihtsustatud on PHP, MySQL ja Apache seadete muutmine. Veebiserveri puhul erilisi arvestavaid alternatiive pole peale IIS, mis on Microsoft'i poolt arendatud veebiserveri tarkvara. IIS on ühilduv ainult Windows'i operatsioonisüsteemidega, universaalsuse huvides Apache on parem valik, sest toetab rohkemaid operatsioonisüsteeme kui IIS.

Andmekaeve programmid ehk rakenduse alamrakendused(vt. alapeatükk 4.3) töötavad serveris, sellepärast on need kirjutatud PHP's, mis teatavasti on *server-side* programmeerimiskeel ning andmed salvestatakse relatsioonilisse MySQL andmebaasi. Andmekaeve on pikaajaline ning mälumahukas töö, sellepärast tuleb kindlasti pöörata tähelepanu PHP konfiguratsioonile ning seadistada kõik vastavalt vajadusele. Praegusel juhul tõstsime lubatavat mälumahtu vastavalt ülesandele ning võtsime ajalised piirangud maha.

Kasutajapoolne rakendus sai kirjutatud kasutades enam-levinuid veebikeeli. Kujundamisel kasutati HTML ja CSS. Veebilehe interaktiivsus ja dünaamilisus loodi Javascript'i, jQuery ning AJAX'i abil. jQuery puhul on tegemist Javascript'i teegiga, mis hõlbustab HTML sisu läbikäia, käsitleda sündmusi, luua animatsioone ning luua AJAX't kasutades side serveriga. AJAX'i(Asynchronous JavaScript and XML) puhul on tegemist veebiarendamise tehnikaga, mille tulemusena saab kliendipoolne rakendus ilma vahelesegamata andmeid serverist. Rakenduse poolt hinnatud kommentaaride polaarsuste suhet kujutatakse Google *Chart Tools* abiga.

4.2 Andmemudel

Andmebaasi tehnoloogia on vabavaraline MySQL. Tegu on relatsioonilise andmebaasiga, mida omab suurfirma nimega Oracle.



Joonis 3: Rakenduse andmebaasi mudel

Alapeatükis 4.3.2 kirjeldatava andmekaeve meetodi puhul salvestatakse andmed *sentiments* andmebaasi *sentiments* tabelisse. Andmed salvestatakse osaliselt **Joonisel 3** kujutatud andmemudeli alusel, vaid osaliselt. Ise roomatud andmed salvestatakse kujul *id, link, author, time, text, vote, active, timestamp*.

- *Id* – sentimentaalset väärtust omava kommentaari identifikaator.
- *Link* - puhul on tegu aadressiga, kust kommentaar on võetud.

- *Author* – kommentaari autor.
- *Time* – aeg, millal on kommentaar kirjutatud.
- *Text* – kommentaar.
- *Vote* – tegu on arvulise väärtusega, mille abil tulevikus saab hinnata rakenduse poolt tehtud hinnangu õigeks pidavust. See väärtus arvutatakse rakenduses inimeste poolt tehtud hääletustulemuste alusel. Oletame, et on mingisugune hinnatud kommentaar, siis inimesel on võimalik anda sellele kommentaarile hinnang, et kas tegu oli korrektselt hinnatud väärtusega või mitte. Inimese poolt tehtud hinnang salvestatakse *vote* tabelisse.
- *Active* – tegu on väljaga, mis näitab kuvamisskriptile, et kas see väli on aktiivseks kasutamiseks lubatud või mitte. Oletame, et *vote* välja näitaja annab märku, et hindaja pole olnud täpne oma hinnangu andmisel, siis see *active* väli muudetakse vääraks ehk pole kasutamiseks lubatud.
- *Timestamp* – ajahetk, mil on kommentaar salvestatud andmebaasi.

Alapeatükis 4.3.1 kirjeldatav andmekaeve meetodi puhul salvestatakse andmed *sentiment* andmebaasi *sentiments* tabelisse. Erinevalt 4.3.2 meetodi puhul, praeguses meetodis salvestatakse andmed täielikult vastavalt **Joonisel 3** kujutatud andmemudelile. Võrreldes 4.3.2 meetodile lisatakse väärtused väljadele *article_id* ja URL.

- *Article_id* – puhul on tegemist artikli identifikaatoriga *ekktt_article* tabelis.
- *URL* – roomatud artikli aadress. Erinevus *link*'i suhtes seisneb selles, et *link* annab täpse URL'i kust kommentaar võeti, aga *URL* annab artikli aadressi, mitte kommentaariumi.

Alapeatükis 4.3.1 kirjeldatav meetod hangib informatsiooni *ekktt* nimeliselt andmebaasist, kus asetseb tabel nimega *ekktt_article*. Seal on meile olulised väljad *id*, *channel_id* ja *URL*.

- *Id* – artikli identifikaator.
- *Channel_id* – väljaanne, kust artikkel võetud.
- *URL* – artikli täpne aadress.

Kui on võetud artikkel koos oma andmetega *ekktt_article* tabelist, siis alustatakse temaga tööd nagu punktis 4.3.1 kirjeldatud. Selle tulemusena määratakse andmebaasi *sentiment* tabelisse *crawled* märge, et see artikkel on roomatud kujul: *id, article_id, URL, timestamp.*

- *Id* – roomatud sündmuse identifikaator.
- *Article_id* – identifikaator *ekktt_article* tabelis ehk artikkel, mille kommentaare on kogutud.
- *URL* – artikli aadress.
- *Timestamp* – ajaline märge, et kuna artikli kommentaaride kogumist alustati.

Hindamise jaoks on *sentiment* andmebaasis olemas *dictonary* tabel, mis omab Tartu Ülikooli arvutilingvistika uurimisrühma poolt loodud tasakaalus korpuse sõnu sageduse järjekorras(kahanevas).

- *Id* – sõna järjekorra number, identifikaator.

- *Frequency* – sõna sagedus korpuses.
- *Word* – sõna.
- *Positive – boolean* väärtus, kas sõna on positiivne või mitte.
- *Negative – boolean* väärtus, kas sõna on negatiivne või mitte.
- *Answered – boolean* väärtus, kas sõna on hinnatud või mitte.

Kui rakendus on hinnanud mingisuguse päringu tagajärjel kommentaare ning lahterdanud nad, kas positiivseteks, negatiivseteks või neutraalseteks, siis inimesele kuvatakse sektordiagramm päritud isiku või brändi kohta. Pärast saab inimene vaadata hinnatud kommentaare ning hinnata nende lahterdamise õigsust. Need hinnangud on andmebaasi salvestatud *sentiment* andmebaasi *vote* tabelisse.

- *Id* – idendifikaator.
- *Sentiments_id* – rakenduse poolt hinnatud kommentaari *id*.
- *Article_id* – rakenduse poolt hinnatud kommentaari artikli *id*.
- *Correct – boolean*, kas rakenduse poolt tehtud hinnang on tõene kommentaari suhtes.
- *Timestamp* – aeg, kuna hinnang tehtud.

Andmebaasi *charset* peab olema *UTF-8*(MySQL puhul *UTF8-general-ci*).

4.3 Sotsiaalmeedia ehk rakenduse informatsiooni päritolu

Käesoleva bakalaurusetöö praktilises osas kasutan sotsiaalmeediast informatsiooni kogumiseks veebiroomajat(vt. alapeatükk 4.3.2) ning DeepWeb'i andmebaasi, mida jooksvalt teise projekti raames veebiroomajad täiendavad. DeepWeb'is asuvad ajalehtede artiklite veebiaadressid. Veebiaadresside abil korjatakse vastavate artiklite kommentaariumitest inimeste poolt kirjutatud kommentaare(vt. alapeatükk 4.3.1).

Sotsiaalmeedia koosneb veebi- ja mobiilipõhistest tehnoloogiatest, mis võimaldavad inimestel, organisatsioonidel ja kommuunidel omavahel suhelda ning arendada reaalajas dialooge. Üks sotsiaalmeedia väljundeid on uudiste- ehk meediaportaalid, kus inimesed saavad kommenteerida päevakorras olevaid probleeme ja arendada arutelu.

Praeguse rakenduse puhul keskenduti eesti keelsetele veebi lehekülgedele. Peamisteks allikateks on Postimehe, Delfi ja Äripäeva portaalid. Delfi alla kuuvad ka Eesti Päevaleht, Eesti Ekspress, Maaleht ja teised Delfi kontserni kuuluvad väljaanded.

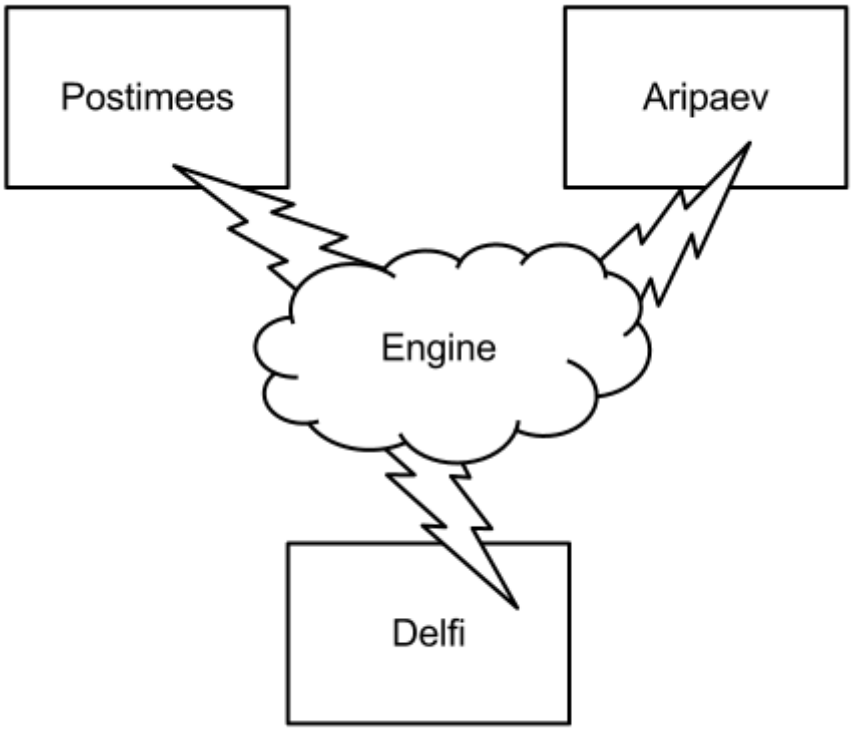
4.3.1 DeepWeb ehk eelnevalt roomatud artiklite kommentaaride kogumine

Teine andmekaeve meetod seisneb selles, et kuna Postimehe ja Äripäeva portaalid pole ilma Javascript'i ohjeldada suutva veebiroomajata efektiivselt roomatavad, siis nende lehekülgedelt andmete kogumiseks kasutame kolmanda osapoole poolt kogutud artiklite

andmebaasi. Artiklite roomamiseks vajaminev informatsioon pärineb deepweb.ut.ee roomajalt.

Kommentaariumite kogumiseks on loodud alamrakendus, mis tegeleb DeepWeb artiklite sorteerimisega ning õigetele klassidele ettesöötmisega. Klassides kogutakse veebilehe spetsiifiliselt kommentaare arvestades nende HTML eripärasusi. Kaevandatud artiklid lisatakse andmebaasi kui „roomatud“. Tulevikus alamrakendus teab, milliseid külastada, milliseid mitte.

Klasse on kolm: Postimees, Aripaev(Äripäev) ning Delfi. Igaüks tegeleb vastavalt omanimeliste URL'dega, kuid Delfiga on asjaolud teisiti. Kuna Delfi kontserni veebilehed on loodud ühe süsteemi peale, siis Delfi klass on võimeline töötlema nii Eesti Päevalehe, Eesti Ekspressi, Eesti Maalehe ja muid Delfi kontserni kuuluvate väljaannete artikleid.



Joonis 4: Graafiline illustratsioon sorteerimisest

Engine.php fail tegeleb *ekktt*(vt. **Joonis 3**) andmebaasist *ekktt_article* tabelist artiklite võtmisega ning *sentiment* andmebaasis vaatab *crawled* tabelist, et kas sealt artiklilt on juba kommentaare kogutud. Kui vastav viide puudub, siis see lisatakse *crawled* tabelisse ning artikkel antakse *engine.php* poolt edasi vastavale klassile, mis oskab seda töödelda. Vastava klassi juurde suunatakse artikkel *ekktt_article* välja *channel_id* järgi.

Kui klassid oma töö tagajärjel leiavad, et vastava artikli kommentaarium omab kommentaare, siis need kommentaarid salvestatakse andmebaasi *sentiment* tabelisse *sentiments*.

4.3.2 Iseseisvalt veebilehtedel roomamine

Delfi kontserni kuuluvate portaalide andmekaeveks kirjutati lihtne veebiroomaja, mille ülesandeks oli etteantud sügavuse ja URL'i alusel roomata veebilehti. Roomajatega võib olla selline probleem, et kui neile anda liiga suur sügavus, siis nad võivad liikuda etteantud leheküljelt ära ning raisata masina jõudlust ning aega tarbetu tegevuse peale. Selle jaoks loodi triviaalne piiraja, milles määrati ära, mis stringi peab URL sisaldama, et URL oleks sobiv. Selline lähenemine piiras jällegi kasutamismugavust. Enam polnud võimalik Delfit ja tema väike väljaandeid roomata ühe sessiooniga.

Ainult Delfi kontserni jaoks loodi veebiroomaja põhjusel, et nende veebilehed ei oma nii palju Javascript'i kui näiteks Äripäev ja Postimees. Viimaste puhul on suurelt jaolt kogu sisu genereeritud Javascripti toel. Roomajad, mis on võimelised jooksutama Javascript'i on praeguse projekti skoobist väljas.

Detailne kirjeldus:

(1) Kasutaja valib portaali, mida ta soovib roomata, määrab sügavuse ning soovi korral sisestab piirava stringi.
(2) Kui kasutaja vajutab roomamisrakenduses alustamisnupule, siis algab etteantud URL'l töö.
(3) Roomaja töötab rekursiivsel põhimõttel, sedasi ei pea tegema külastatavate URL'dega listi. Selline lähenemine omab ka nimetust süvitisotsing ehk kõige pealt minnakse ühtede URL'de pidi maksimaalse sügavuseni ning kaevatakse vajaminevad andmed.
(4) Roomaja loob alguses staatilise listi, kuhu pannakse kirja külastatud URL'd.
(5) Roomaja kontrollib seejärel, kas seda URL'i on varem roomatud, kui jah, siis katkestab töö kui mitte, siis jätkab. Praeguse alamrakenduse puhul lisati veel selliseid piirajaid, et kui URL sisaldas vene keelsele leheküljele viidet, siis katkestati samuti töö, kuna peamine eesmärk on luua eesti keelne rakendus.
(6) Igal leheküljel kontrollitakse teatuid HTML klasside olemasolu, milles võib peituda meile vajalik informatsioon. Olemasolu korral viiakse ellu andmete kogumine ja andmebaasi salvestamine(vt. **Joonis 3**). Andmed säilitatakse *sentiment* andmebaasi ja *sentiments* tabelisse.
(7) Viimasena võetakse leheküljel olevad URL'd ning kutsutakse roomamisfunktsioon välja värskete URL'dega välja.

4.4 Sentimentaalsuse hindamine rakenduses

4.4.1 *„Bag of words“* ehk „Kott sõnadega“

„Kott sõnadega“ mudel on oletus keele masinõppes ja infootsingus, kus „sõnahunnikute“ põhjal algoritmid määravad mingisuguse teksti sisu oleku. Olekute all peame silmas, et kas see uuritav sisu on meile vastuvõetav(otsingumootori põhjal, kas vaadeldav dokument on sobiv või mitte), näitena sentimentaalsest valdkonnast, kas tegu on positiivse sisuga või negatiivse sisuga tekst. Selles meetodis on tekst esitatud järjestamata sõnade kogumina, kus ei arvestada grammatikat ega sõnade järjekorda. Kui tekst sõnade kogumina, siis võrreldakse kogumit ühe või mitme koti sisuga. Olenevalt püstitatud eesmärkidest ja „kottide“ omadustest võtavad algoritmid vastu otsuseid. Kasutatakse dokumentide klassifitseerimisel, sentimentaalsuse hindamisel, filter-rakenduste puhul.

Näiteks *e-mail* rämpsposti filtrite puhul kasutatakse kahte erinevat kotti sõnadega. Üks kott omab spämmkirjadele omaseid sõnu ning teine korralikes kirjades kasutatavaid sõnu. E-kirja hindamisel tekitatakse kirjast hunnik sõnu ning võrreldakse sõnu kottide sisuga. Pärast filter-algoritm otsustab, et kas on tegu rämpskirjaga või mitte.

4.4.2 Sentimentaalsuse hindamise meetod

Meid huvitab sõnade polaarsus, kas nad on positiivsed või negatiivsed. Neutraalseid sõnu me silmas ei pea. Sõnade semantilise orientatsiooni määrame käsitsi kasutades „kott sõnadega“ meetodit. Loome polaarsusleksikoni, mis omab kahte „kotti“ – positiivseid ja negatiivseid sõnu. Polaarsusleksikon luuakse kasutades Tartu Ülikooli arvutilingvistika uurimisrühma poolt loodud sagedustabeli alusel, mis on loodud tasakaalustatud korpust kasutades [16].

Hindamisel võtsime aluseks Hatzivassiloglou ja McKeown’i uurimustöös saadud tulemuse, kus 657 posiitvsete omadussõnadega ning 679 negatiivsete omadussõnadega saadi hindamisel esimesel korral 89% täpsust ja teisel korral 97% täpsust [17]. Käesolevas töös me ei eeldagi, et meie saame sarnased arvulised tulemused hindamisel, kuna hindamise algoritm erineb. Hatzivassiloglou ja McKeown’i uurimustöös olnud positiivsete ja negatiivsete sõnade arvu võtame endale eeskujuks, et selgitada välja, kui palju peaks see vähemalt olema polaarsusleksikonis. Rakenduses hindasime ära 1613 sõna, millel esinemissagedus korpuses vähemalt 50. Positiivseid 663 ja negatiivseid 950. Hindamisel ei tehtud vahet sõnaliikidel ehk ei spetsialiseeritud kindlatele, sest eesti keeles pole veel sõnaliikide mõju sentimentaalsusele hinnatud.

Loodud polaarsusleksikon kahe „kott sõnadega“ puhul peame meeles, et tegu on manuaalselt loodud leksikoniga ning sõnadele antud hinnangut ei saa pidada objektiivseks, sest hindaja puhul võib mõndade sõnade puhul olla mõjutatud arvamus ehk subjektiivne, mis võib mõne teise isiku omaga erineda. Hinnanguid andes lähtuti põhimõttest sellest,

kuidas need sõnad omavad polaarset tähendust lähtuvalt hindaja poolt läbi mängitud erinevatest kontekstidest.

4.4.3 Sentimentaalsuse hindamise rakendus

Rakenduses kasutasime phpInsight programmi [18], mille puhul on tegemist inglise keelse sentimentaalsuse hindajaga. Tegu on vabavara ning avatud koodiga arendusega, mis baseerub naiivsel Bayes'i klassifitseerijal, mis kujutab endast lihtsat tõenäosuslikku klassifikaatorit, kus kasutatakse Bayes'i teoreemi koos tugevalt(naiivsete) sõltumatute eeldustega .

Esmane probleem phpInsight programmiga integreerimisel käesoleva lõputöö rakendusega seisnes selles, et programm on mõeldud inglise keelsete tekstide hindamiseks. Sellepärast modifitseerisime programmi eemaldades ebavajalikud inglise keelsed listid. Loetelu muudatustest:

(1) Eemaldasime phpInsight'st inglise keelsed polaarsusväärtused ning polaarsusleksikonist neutraalsuse klassi;
(2) eemaldasime ignoreeritavate sõnade nimekirja;
(3) eemaldasime eesliidete nimekirja;
(4) kirjutasime ümber stringi puhastamise funktsiooni, et puhastada sisendina antud string üleliigsetest karakteritest.

Ühendasime phpInsight'i *sentiment.class.php* olemasoleva andmebaasiga ning lõime klassi polaarsusleksikoni, mis koosnes positiivsetest ja negatiivsetest sõnadest, mis eelnevalt manuaalselt hinnatud. Lisaks lõime funktsiooni, mis eemaldas lausest kõik ebavajalikud märgid, kuni alles jäid ainult sõnad. Lõpuks jäi lausest alles ainult hunnik sõnu, mida pärast võrreldi leksikonis olevate sõnadega.

Kui phpInsight oli eesti keelseks kasutuseks kõlblik, siis rakenduses loodi klass S*entiment* , kus pärast kasutasime hindamisel funktsiooni *categorize*, mille sisendiks sotsiaalmeediast pärit artikli kommentaar ehk string. Kui ette antud string sai hinnatud, siis tagastati kommentaari polaarsus(positiivne, negatiivne, neturaalne) hindamisfunktsiooni väljakutsunud programmijupile.

4.5 Veebirakendus

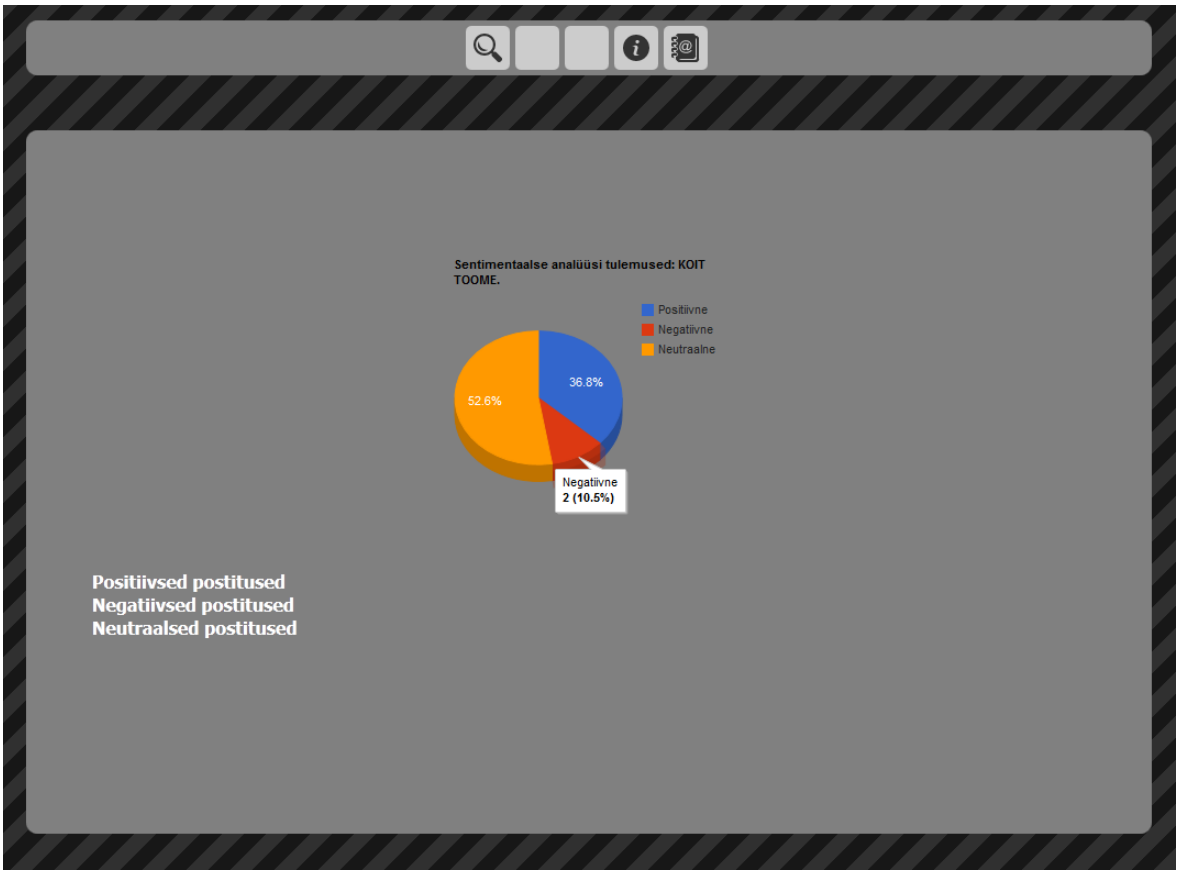
Veebirakenduse puhul on tegemist kliendimasinas baseeruva rakendusega, mille ülesandeks on olla graafiliseks lüliks kliendi ning serveri-poolse rakenduse vahel. Käesoleva projekti veebirakendus annab kasutajale võimaluse teha päringuid ning saada vastavate päringute kohta tagasisidet, milleks on hinnatud kommentaarid.

Joonisel 5 näeme veebirakenduse esilehte ning pooleli olevat toimingut, kus on näha, et praegusel hetkel on kasutaja sisestanud otsinguvälja Eesti muusiku Koit Toome nime.

Joonis 5: rakenduse esileht, pooleli olev toiming.

Kui kasutaja on vajutanud „Otsi“ nupule, siis sellega alustatakse suhtlust serveriga, kasutades AJAX’t. Kasutajal ei laeta kordagi uuesti veebilehte, kõik uuendused luuakse jQuery abil eemaldades või lisades HTML ja CSS klasse. Otsingu nupule vajutades alustas server vastava otsingusõna järgi andmebaasist kommentaaride otsimist ning nende sentimentaalsuse hindamist. Sisestatud otsingusõna leitakse andmebaasist viisil, kus otsingusõna pannakse suurtähtedesse, samal ajal iga kommentaar pannakse ka suurtähtede-formaati. Kommentaar osutub valituks juhul kui kommentaari string omab endas täpselt otsingusõna. Et kasutaja saaks aru, et server on töös, siis ekraanil töötab punane *gif* tüüpi animatsioon.

Kui server on saanud valmis kommentaaride kogumise ja hindamisega saadetakse *JSON*’i abil vajaminev andmestik kliendi masinasse, kus kasutades Google *Chat Tools*’i abil kuvatakse sektordiagrammi abil polaarsuste suhte(vt. **Joonis 6**). *JSON*’i puhul on tegemist kergekaalulise andmevahetusformaadiga.



Joonis 6: andmete kuvamine.

Andmete vaatamise järel on võimalik kasutajal uurida erinevaid sotsiaalmeedia kommentaare vastavalt nende liigitusele. **Joonis 6** on näha, et all vasakul nurgas on valges kirjas kolm erinvat fraasi. Nende fraaside puhul on tegemist linkidega, mille all saab näha nagu eelnevalt mainitud automaatsel viisil hinnatud kommentaare. Lisaks vaatamisele on võimalik kasutajal hinnata tekstide sentimentaalsuse õigsust. **Joonis 7** on näha, et Koit Toome päringule vastuseks saadud positiivsed kommentaarid on hindamiseks kuvatud ning esimene kommentaar on kasutaja poolt juba hinnatud, see märgiti linnukesega rohelisel taustal. Kohati on teatud kommentaare tuhandeid ning seetõttu liigutatakse veebilehtede vahel informatsiooni PHP sessioonide abil. PHP sessionite puhul on tegemist globaalsete muutujatega, mis loomisel on igale kliendimasinale universaalne ehk pole vaja karta, et server kuvab andmeid valele kasutajale.

Joonis 7: rakenduse poolt hinnatud kommentaaride õigsuse hindamine.

4.6 Kasutamisjuhend

Installeerimine:

(1) Paigaldada infrastruktuur: Apache veebiserver, MySQL server, PHP. Soovitavalt WAMP(Windows-Apache-MySQL-PHP) või LAMP(Linux-Apache-MySQL-PHP) olenevalt operatsioonisüsteemist.
(2) Rakenduse andmebaasi mudel paigutada MySQL serverisse *sentiment* nimelisse andmebaasi:
    1. Lood *sentiment* andmebaasi;
    2. *sentiment* andmebaasi saad vajaliku andmemudeli kui jooksutad *sentiment.sql* faili;
    3. järgmiseks tuleb andmebaasi *dictionary* tabelisse polaarsusleksikoni andmed ehk jooksuta *dictionary.sql*;
    4. testandmete saamiseks lisa andmebaasi *test.sql*, mis lisab andmebaasi 25 000 kommentaari kirjet.

(3) *Config.php* failis ära märkida andmebaaside ligipääsuks olevad andmed. Eeldab, kas ligipääsu *ekktt* andmebaasi või *ekktt* andmebaasi artiklite väljavõtet. Väljavõtte puhul lisada artiklid oma MySQL serverisse *ekktt* andmebaasi:
    1. Loo *ekktt* andmebaas;
    2. jooksuta *ekktt_database.sql* fail ning selle tulemusena luuakse tabel, mis omab kirjeid.
(4) Rakenduse failid WAMP/LAMP puhul pakkida lahti *www* kausta. Kui WAMP/LAMP tehnoloogiat ei kasuta, siis pakkida kaust lahti Apache veebiserveri poolt hallatavasse kausta, millele pääseb brauseriga ligi.

Kasutamisjuhend:

(1) Andmekaeveks jooksutada, kas *crawler* kaustas *engine.php* või *crawler.php*. Mõlemal juhul eeldatakse, et andmebaaside(*sentiment* ja *ekktt*) olemasolu ning *config.php* olevate andmete olemasolu ja õigsust.
(2) Kui andmekaeve on alanud ning esimesed kommentaaride kirjed andmebaasi tekkinud, siis on võimalik kasutada veebirakendust, kus saab teha päringuid ning vastavate otsitavate parameetrite olemasolul tagasiside ehk luuakse sektordiagramm hinnatud kommentaaridega.

5 Eksperimendid

5.1 Hindaja

Rakenduse sentimentaalsuse hindaja täpsuse testimiseks lõime sajast kommentaarist koosneva faili, kus kommentaarid olid automaatse hindaja poolt ära hinnatud. Pärast automaatset hindamist hinnati iga kommentaari kohta antut hinnangut, kas automatiseeritud rakenduse poolt antud hinnang on tõene või väär.

Testandmete valimine andmebaasist tehti automaatiseeritul viisil. Valiti 100 erineva andmebaasi identifikaatoriga kirjet, mis olid väiksemad või võrdsed 100 sõnaga. Saadud kirjete kommentaarid lasti läbi sentimentaalsuse hindaja, mis määras stringi polaarsuse(positiivne, neutraalne või negatiivne). Automatiseeritult kogutud testandmed salvestati faili ning lugejale lisatud lisasse lugemiseks ja uurimiseks(vt. Lisa 1). Järgmise käiguna hinnati phpInsight hindaja korrektsust kahe isiku poolt.

		Automaatne hinnang		
Inim-hinnang		Positiivne	Neutraalne	Negatiivne
	Positiivne	9	3	0
	Neutraalne	3	55	6
	Negatiivne	5	8	11

Tabel 1: Automatiseeritud hindaja tulemus 100 juhuslikult valitud kommentaari kohta

Tabelis 1 on näha tulemusi, kus 75 kirjet sajast võib lugeda inimhinnangul õigeks ehk 75% täpsus. Kindlasti tuleb ära märkida fakt, et inimhinnang on subjektiivne, mitte objektiivne. Katses hindasid automatiseeritul viisil saadud hinnanguid kaks erinevat isikut, kelle hinnangud muudeti läbi diskussiooni ja kompromisside üheks.

Positiivseid kommentaare oli automaatse hinnangu puhul 17, kuid inimhinnang pidas õigeks neist 9. Mis annab positiivsete kommentaaride hindamise täpsuseks siin katses 53%. Negatiivsete kommentaaride puhul leidis masin, et 17 kommentaari on negatiivsed, kuid inimhinnang alusel ainult 11. See annab negatiivsete kommentaaride täpsuseks 65%. Neutraalseteks pidas automaatne hindaja 66 kommentaari, inimhinnang kinnitas 55 korrektsuse, mis annab täpsuseks 83%.

5.2 Andmekaeve klasside ja hindaja kiirus

Teises eksperimendis hindasime erinevate andmekaeve mooduli kiiruseid, välja selgitamaks, kui kiiresti jõuavad erinevad klassid 4.3.1 meetodis artiklite kommentaare andmebaasi salvestada. Eesmärgiks on näha, et kui kaua kulub kommentaaride kogumine portaalidest. Postimehe ja Delfi klasside puhul võeti hindamiseks artiklid, millel on kommentaare vahemikus 50-150. Äripäeva puhul võeti 30-60, sest seal ei leidu eriti artikleid, mis omaksid sadades kommentaare.

Oluline on ära märkida, et testitava masinal on ligipääs Internetti võimaldatud ühendusega, mille kiirus on 1Mbit/s. See on üks tähtsamaid tegureid, miks kogumine võtab nii kaua aega.

KLASS	1	2	3	4	5	6	7	8	9	10	11	12
Postimees	6	10	9	9	10	9	4	7	8	7	8	7
Delfi	9	7	7	7	7	7	4	4	3	4	4	4
Äripev	11	11	10	11	10	12	13	14	16	14	14	15

Tabel 2: Klasside kiirused andmekaeve meetodis 4.3.1, tabeli andmed on sekundites

Postimees mõõtmiste 1-6 puhul omas artikkel 102 kommentaari, 7-12 puhul 57 kommentaari.

Delfi mõõtmiste 1-6 puhul omas artikkel 115 kommentaari, 7-12 puhul 52 kommentaari.

Äripäev mõõtmiste 1-6 puhul omas artikkel 35 kommentaari, 7-12 puhul 51 kommentaari.

Liikudes ringi Äripäev kodulehel, on tunda, et serveri reageerimisaeg on tunduvalt aeglasem võrreldes teiste uudisteportaalidega. Kohati on aeglus isegi kasutajat ärritav.

Tabelis 2 on märgitud klassid, kus 1-6 puhul kasutati ühte artiklit ning 7-12 teist artiklit. Tabelis on märgitud klassidel kulunud aeg sekundites.

Kasutajamugavuse säilitamiseks on oluline, et hindamine toimub võimsa protsessori peal(arenduses kasutati Intel'i *i5 760* neljatuumalist protsessorit ning katsetati ka Intel *Core Duo* protsessorit T7100), sest vastasel juhul läheb hindamise aeg liialt pikaks ning pole enam kasutajatele aksepteeritav tulemuse ootamiseks kuluv aeg. Hindamisel kulub aega andmebaasist hindajale ressursside otsimine ning hindamise protsess ise. Näiteks sisestamisel otsingusõna „Venemaa", annab andmebaas ja hindaja 2117 positiivset, 2811 negatiivset ja 4446 neutraalset väärtust ehk hindaja hindas 9374 kommentaari ning seda 10 sekundi jooksul. Otsingusõna „Ansip" puhul oli vasteks 6693 kommentaari ning sellise hulga hindamiseks kulus 8 sekundit.

6 Tulemus

Projekti alguses seatud eesmärgid rakendusele said täidetud ehk automatiseeritud andmekaeve ning triviaalne hindamisklass töötavad. Andmebaasi täiendatakse regulaarselt nii kaua kui DeepWeb andmebaas omab artikleid, mida pole veel rakenduse poolt töödeldud. Andmete kaevet on pidevalt muudetud ning jõutud etappi, kus artiklite kommentaaride kogumine on aksepteeritava kiirusega ehk loetud sekundite jooksul võetakse kommentaarid ühe artikli juurest ära. Postimees.ee puhul oli probleemiks, et Postimehe veebileht omab liialt palju linke ning nende töötlemine võtab liialt palju aega(nõrgematel masinatel võis võtta minuteid), seetõttu sai arendatud alternatiiv, mille tulemusena saadi aksepteeritav tööaeg. Lahendus seisnes sellest, et loobuti linkide kaudu kommentaariumi lisalehekülgede otsimisest.

Modifitseeritud phpInsight poolt tehtavat hinnangut kommentaarile saab rakenduses hinnata, et lõpptulemusena saavad arendajad jälgida hindaja või polaarsusleksikoni valupunkte. Kasutajale on antud võimalus anda tagasisidet reaalajas vastavalt kommentaaridele.

7 Kokkuvõte

Bakalaureusetöö raames loodi rakendus, mis on võimeline Eesti sotsiaalmeediast koguma artiklite kommentaare ning salvestama andmebaasi. Andmete kogumiseks on loodud kaks erinevat alamrakendust: esimese puhul kasutatakse DeepWeb andmebaasi kommentaaride kogumiseks ja teisel juhul omab rakendus algelist roomajat, mis kogub ressursse. Esmane eesmärk oli andmeid koguda ning seda automatiseeritult, teiseks tuli luua hindaja. Hindaja pidi andma hinnangu talle sisendina sisestatud tekstile. Hindaja loodi phpInsight programmi ümberkirjutamisel, sest esialgne versioon oli mõeldud inglise keelsete tekstide hindamiseks. Kolmandaks loodi veebirakendus, mis võimaldab teenuse kasutajatel teha päringuid ning saada Eesti sotsiaalmeediast tagasidet.

Töös toodi välja sentimentaalse analüüsi ning veebiroomajate tagamaad, andes lugejatele elementaarne arusaam temaatikast. Kirjeldati rakenduses kasutatavaid tehnoloogiaid, loodud andmebaasi mudelit, andmekaeve protsessi ja hindaja loomist ning selle tööd. Lisakes näidati eksperimendi raames hindaja täpsust.

Tekitades võrdluse tasuliste teenustega on käesoleva bakalaureusetöö raames loodud rakendus suhteliselt triviaalne, seevastu tasuta pakutavatega võrdväärne. Rakenduse muudab eriliseks asjaolu, et teadaolevalt puudub hetkel mõni teine rakendus, mis koguks eestikeelseid sotsiaalmeedia tekste ning hindaks nende semantilist orientatsiooni ja lõpuks annaks kasutajale tagasisidet. Kindlasti ei saa rakendust pidada Eestis sentimentaalse analüüsi valdkonnas läbilöögiks, vaid pigem katseks luua midagi töötavat ning uurida potentsiaalse magistritöö jaoks võimalikke valupunkte.

Abstract

Sentiment Analysis of Estonian Mainstream Media Article Comments

Bachelor's thesis

Siim-Toomas Marran

With the growth of different social media channels like social network services, blogs, forums, microblogs and so on, there has been huge increase of opinion-rich resources. Opinion-rich resources contain valuable information for the different government organisations, educational institutes and private sector companies. By the opinion-rich resources we mean the information with semantic orientation or in another terms polarity. Polarity can be described with values *positive*, *neutral* and *negative*. The huge amount of different social channels brings us to the point where we need automated info-seeking and gathering programs which collect opinion-rich resources. Since social media possesses billions entries of data, it is unreasonable to process information by people manually. This situation creates a need for different natural language processing, computational linguistics and text analytics techniques to identify and extract subject information in source materials. The whole process which has been previously described is called sentiment analysis or also opinion mining in some scientific communities.

This thesis paper concentrates on the data mining from the Estonian online media portals and evaluating collected resources with trivial „bag of words" technique As whole the entire process is called opinion mining. Data mining took place in Postimees, Delfi, Eesti Päevaleht, Eesti Ekspress and Äripäev sites. „Bag of words" technique uses polarity dictionaries which contains lists of positive and negative words. In this project polarity dictonary was created manually from the frequency list where we evaluated words as postive or negative. When the application had opinion-rich resources from the media and the polarity dictionary, we created with the help of phpInsight(English sentiment evaluator) Estonian sentiment evaluator, which now made possible the web application to serve users. Data will be shown to the users with a web application where the users have possibility to create queries about different individuals and brands in which they have an interest in. Comments' polarity data about the interest will be shown for the users in pie chart form. Also the users have a possibility to judge Estonian sentiment evaluator's rightness. In the paper every important aspect is described. Also there is an experiment to judge the accuracy of the evaluator and the data mining speed on the various sites.

In conclusion, in this bachelor's thesis we have accomplished raised goals by constructing, to our knowledge, the first Estonian opinion mining solution for Estonian social media.

Even though it is the protoype, it shows out various problems about the sentiment analysis and gives good feedback for potential master's thesis.

Kasutatud kirjandus

[1] Wikipedia. Facebook.
http://en.wikipedia.org/wiki/Facebook.

[2] Wikipedia. Twitter.
http://en.wikipedia.org/wiki/Twitter.

[3] Wilson T, Wiebe J, Hoffman P. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. 2005.

[4] Wikipedia. Sentiment analysis.
http://en.wikipedia.org/wiki/Sentiment_analysis.

[5] Turney PD, Littman ML. Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus. 2002.

[6] Bang B, Lee L. Opinion mining and sentiment analysis. Pre-publication version; 2008.

[7] Radian6.
http://www.radian6.com/.

[8] Lithium.
http://www.lithium.com/.

[9] MediaVantage. http://www.mediavantage.com/.

[10] Socialmention*.
http://www.socialmention.com/.

[11] Sentiment140.
http://www.sentiment140.com.

[12] Thelwall M. A web crawler design for data mining.

[13] Castillo C. Effective Web Crawling. 2004.

[14] Gulli A, Signorini A. "The indexable web is more than 11.5 billion pages". 2005.

[15] Wikipedia. Apache HTTP Server.
http://en.wikipedia.org/wiki/Apache_HTTP_Server.

[16] Eesti Kirjakeele Sagedussõnastik.
http://www.cl.ut.ee/ressursid/sagedused1/index.php?lang=et.

[17] Hatzivassiloglou V, McKeown KR. Predicting the Semantic Orientation of Adjectives.

[18] Hennessy J. phpInsight. https://github.com/JWHennessey/phpInsight.

Lisad

Lisa 1 :hindamise testandmestik

Nr	Kommentaar	Automaatne hinnang	Tõene
1	inglise tiblasid oli vähe enamus olid nõukogude tiblad	neu	1
2	Selliste ebaõiglaste kohtuotsuste tõttu on väga raske usaldada ja armastada eesti riiki. Kohtunike riigiesindajate vastutustundetu otsuste tegemine peegeldab meie riigi jõhkrust ja ükskõiksust kannatanute vastu. Selle tüdruku elu on hävitatud pervert on saaanud õiguse ja vabaduse uusi ohvreid või sama tüdrukut ka edaspidi ära kasutada. See on märk kogu eluks kuid mitte ületamatu. Tüdrukule julgust ja õnne edaspidiseks eluks.	neg	1
3	Eiselle all on silmas peetud norrakaid kes ohverdasid oma elu Eestit kaitstes.	neu	1
4	kahjuks kuna põeb aidsi ei karista teda vangikongis ka keegi	neg	1
5	See artikkel ajas naerma. Nii haledat reklaami poleks osanud oodatagi.Sattusin sinna kuulsasse hello kiti poodi Ülemistes.Oh taevas ehtne Tiimari osakond...ainult hinnad ajasid naerma.Tillukese lapse hommikumantel üle tuhande krooni.Tilulilu padjakesi ja kotikesi..... hinnad ajasid ikka naerma.Kodus tikime mõnusale padjale ise lapse nime ja uinub nii magusasti tudule.Ilusa rääkiva nuku sain natuke eemalt Juku poest. Rõõmu kauaks ja rahakotile ei teinud ka liiga	pos	0, neg
6	... juhtimisõiguseta ja purjus peaga ringi kihutades panni ta ohtu nii enese kui ta ...Mis pann	neg	1
7	Testitud üle iPhone 3GSi EMT 3G leviala. Asukoht Tallinn Kalamaja.Download: 098MbsUpload: 031MbsJa võin kinnitada et siin ei tule EMT 3Gst kunagi rohkem välja. Olen nendega siin küllalt piike murdnud aga kõik mis nad ütlevad tehnik kontrollis kõik korras. Augusti lõpust kolin EMTist minema ka.	neu	1
8	Muide märgistasin toimetuse jaoks ära kaks kommi mis sisaldavad otsest üleskutset tapmisele 1. astme kuriteole.Pärisorjastamise ennetustöö 21.01.2009 04:48Pedofiilide kraaksatused 21.01.2009 09:39Vaatame kui kiiresti toimetus need koristab või kas üldse. Et kas on see leimijutt vaid niisama võbelev õhk või on vaid juudid ja meie suuremad ärimehed need keda ei tohi netis ähvardada.	neu	1
9	Peaasi et iniseda saaks onju.Aga sel korral Sõbraga nõus pronks on teile ju kindlustatud. Esimesed kaks kohta jagavad omavahel niiehknaa Tartu ja Cramo. Seda kumb on näha kevadel. Ei usu et Rapla või TTÜ suudaks Tarva lätlastele jah just lätlastele ilma lätlasteta oleks ilmselt pronksi püüdmine väga raske vastu saada.	neg	1

10	Kruuda ei jõudnud piimaringilt tagasi	neu	1
11	Hamiltoni.	neu	1
12	Kui see täht on 620 valgusaasta kaugusel ja lähitulevikus meid valgustama hakkab siis on ta juba sajandeid tagasi plahvatanud ning alles nüüd jõuab supernoova footonkiirgus meie meeli erutama. LAHE!	pos	1
13	Euro on nende mängumaa nii et kui peaksid eurole vastu hakkama saad ka pommi et siis aralt saba jalge vahel kummardama hakata	neu	0,neg
14	Aga kes on poisi selja taga see hoopis huvitab. Väita. Et Tsahkna teda ei tunne pole eriti usutav.Siis tekkib küsimus mis suhteid on Tsahknal selle selliga see oleks põnev lugeda ja siis saaks ka selgeks miks mehike saab nii laamendada.	pos	0,neu
15	Kus sa kahekordse hinnavahe oled leidnud ehk jagaksid infot teistega ka.	neu	1
16	Tipparstid tegid oma tööd ja seda nii hästi kui suutsid. Arst ei ole jumal.	pos	1
17	Soovitan nn Steliori toidulisandeidSellega on kõik öeldud!Õppida kaugel ja tulla kodumaale tagasi nii lolli ja tõestamatu jutuga võib ainult firma müügiplika!	neu	0,neg
18	Pedofiilidega on lihtne kui lähevad laste kallale ja õnnestub mõni isend neist teolt tabada siis tuleks ta surnuks peksta muidu juhtub jälle see et läheb maksumaksja kulul vangi ja sealt kunagi väljudes tegutseb edasi. Meie väike vabariik pole nii rikas et maksumaksja kulul igasugu perverte nunnutada järjest rohkem hakkab pervertide ja pättide nunnutamine ennast kätte maksma ja kunagi võime lõpetada nagu õudusfilmis et meid valitsevad pättide ja pervertide gängid !	pos	0,neg
19	väga mitte lugupeetav autor tee palun endale selgeks miks suurbritannias sellises koguses neid juhtumeid onpoliitkorrektsuse ühiskonnas läheb statistikasse sisse ka see kui isa oma lapsi mänguväljakul pildistab ja mõni hüsteeriline sinusugune korralik kodanik politsei kutsub.http:yro.slashdot.orgarticle.pl?sid09...kas selline oleks siis korralik ja lapsi kaitsev ühiskond?arenenud riikides pidada täiskasvanud juba lapsi kartma sest lihtsalt vale pilk võib süüdistuse tuua	neg	1
20	Haigekassal on raviteenuste jaoks kindel hinnakiri ja Vähikliinik oleks teenuseid osutanud sama hinnakirja alusel. Kui erakätes Vähikliinik suudab selle hinnakirja juures sama või isegi paremat teenust osutada kui Regionaalhaigla siis on Regionaalhaiglas midagi väga valesti.Maksumaksja jaoks pole vahet kas raha läheb eraettevõtja taskusse või ebaeffektiivse Regionaalhaigla bürokraatia ja juhatuse taskusse.	neu	1
21	Kas Jaapani 9.0 magnituudine maavärin aetakse kah tormide ja maalihete süüks?	neu	1

22	Aiai kallid ehitajadNÜÜD SAATE TE KÕIK NEED HAIGUSED JA HÄDAD OMALEMIS NENDE MÜNTIDE OHVERDAJATEL OLID. Pahapaha lugu. Ja ega Kiudsoogi puutumata jää. Ja kindlasti põdesid paljud ohverdajad ka katku....	neu	0,neg
23	siis nüüd öelda taheti nagu muudegi teemadega muudkui käib üks sama vile aga uut villa pole peale kasvanud.	neu	1
24	inx Sulle paar nõuannet ka. Kui Sa oletad et kõige viimasel juhul oli tegemist kuriteokatsega siis tulnuks pöörduda politseisse. Järve Selveri plats on talutavalt hea videovalve all. Ning kui hästi läheb siis saab need salvestused sealt ka kätte ja sealt näod ja numbrid. Ei oleks enam mingit massihüsteeriat oleks konkreetne kuritegu või selle katse.Aga seda sa ju ei teinud? Isegi võimaliku ohvri nime ei pannud kirja? Lihtsalt solgud siin massihüsteeriaga kaasa ja unistad võimalusest kellegi suhtes kuritegu sooritada tsiteerin: Kahetsen et kasvõi profylaktika mõttes ei pannus neile munadesse..	neu	0,neg
25	vaerske leiva peal mehetegusid teha...	neu	1
26	Kaudsed maksud on reformkommunistliku maailmavaate nurgakivi ja teadlik valik sest see võimaldab nagu muuses rahvast lollitada.KÕIK TE ELATE VEEL TÄNA UIMAS NAGU EESTIS OLEKS MADALAD MAKSUD!!!PALK ON MEIL VÄIKE!	neu	0,neg
27	Ma esitaks pigem retoorilise küsimuse et milline on odavaima hinnaga pakett millega saab omale talutava kiirusega interneti? Starman pakub mingit asja 100kr kuus ja selle kiiruse eest ei saa hiiresgi. Andke 100eeku eest 300500Mbit ja ongi teil tuhandeid kliente juures. 300kr kuus on minu jaoks oluline summa et seda mitte kulutada ja seega kodus internetti ei oma.	neg	1
28	..VÄGA oluline killuke. Ja väikestest killukestest suur asi muide koosnebki!	neu	1
29	Tänan muretsemast meie katusega on endiselt kõik vägagi korras ju siis Aikumeil siin maal kirjutatakse nimed ja lause algus suure tähega on head tööd teinud: Ehk soovite veel kuidagi anonüümses foorumis oma kõrgelaubalisust demonstreerida?	neu	1
30	kes veel ütleb et andres jääb 7900p meheks.mokk maas jah.	neu	1
31	Enamuses EL pole sellist vanemahüvitist olemaski järelikult sellisest direktiivist ei saa seal europarlamendis juttugi olla.	neu	1
32	Ruslan Romanov 06.04.2010 18:59Pole parata need kohalikud võitlevad ateistid ongi siin need kõige häälekamad... ja kahjuks ka kõige sisutühisemad.	neg	1
33	Bush väärib minu silmis sama karistust mis natsiladvik Nürnbergiski.	neu	1
34	Vahepeal kutsub roodu ülem mehed tee algusesse tagasi ja alustatakse otsast peale.Kogu artikli jooksul kasutatud õiget sõna KOMPANII ja siis venekeelne pläkk ROOD.nu PIDUR!	neu	1

35	seetõttu peaks olema nii et korralikult sorteeritud prügi viiakse kokkulepitud! graafiku järgi minema tasuta.Tasu võetakse sorteerimata prügi eest ka sellise prügi eest mille koht on kusagil jäätmejaamas kui inimene ise pole viitsinud viia.Meeletu tasu st. väga suur trahv neile kes prügi kuhu iganes sokutavad!	neu	1
36	pole Kaiale hetkel kõige tähtsam rahanumber,vaid on vaja mänge ja oma aasta alguse hea vorm taastada. Edu talle selleks turniiriks ja soovin,et saaks mängida finaalini välja.	pos	1
37	Tänud! selge pilt eksimine pidi inimlik olema :	neu	0,pos
38	Kurkide marineerimis mooduseid on nii palju ja see mis seal pakil kirjas on on täitsa ehe varjant minagi nii kurke sisse teinud	neu	1
39	Ärge jutustage 177x nädalas vaid rääkige inimeste keelt.Moskva 3x päevasPeterburi 3x päevasKui siis veel vaja eks öelge et nädalavahetusel miski muu asi.Ei ole ju raketiteadus!	neu	1
40	...ja täna ikka poisil puss väljas. Paistab nii. JUba oli ka kuulda et Alpides läheb raskeks ja selles stiilis.	neu	1
41	kuidas ikka sinna jälle on kogutud kokku selline rahvas kes mart laari üle ilkumises oma elu mõtet näevad ? Miks nad on erakonnas mille auesimeheks Mart Laar looedtavasti valitakse ? Kas ilkumine on inimõigus ?	neu	1
42	pole muud kui Ravimiameti bisnise kahju.Patsiendil vahet pole.	neu	1
43	to: w 09.05.2010 12:47Hästi öeldud!	pos	1
44	Iraani jahitakse!Andes rahvale usku et nad võidavad igal juhul oleks see USAle kõige demokraatlikum vallutus võrreldes eelmiste riikide rahuvalvamistega!Eks siis hakkab ka selguma kes juhib seda lombitagust sõjamasinat.	neu	1
45	Rõõmusta pööbel !! Maxuamet juba ootab näppude taskusse pistmist.....	neu	0,pos
46	kas reform maksis siis vähe et õigel ajal õiget numbrit näidata??	neu	1
47	Kena temast et abikaasaga ikka räägib mis sest et juriidikast.	pos	1
48	ikka kokokkutad siis vevenib ikkkka pipikaks mömöla küküll	neu	1
49	Läänes toimib selline süsteem juba pikemat aega. Parklasse sisse ja väljasõidul tehakse autost automaatselt pilt. Arvutisüsteem peilib lepingut rikkunud ehk määratud aja ületanud sõidukid ja juba nädala jooksul saabub täiesti ootamatu trahvinõue.	neu	1
50	Seltsimees Raid millegipärast vaikib tagasihoidlikult nendest pisiasjadest kui palju maksis lõpuks vee hinna kontrolli all hoidmine. Seltsimehel vist plaan kesikuks hakata seepärast sellised poolikud artiklid.	neu	1
51	Ju siis Etol on ainult palk suur siin võetakse ka juurde sponsortehingud.	neu	1
52	Ja kes krt siin kobisevad kommentaariumis??	neu	1

53	heheheeee	neu	0,pos
54	A mis siis saaks kui keegi tõmbaks läbi Harri naise või ta tütre? Mul on põhjust arvata et siis muusika muutuks. Minu jaoks muusika muutuski siis kui ma ise jäin süüta süüga ja ma ei peatu enne kui süüdlane selle oma õlgadele võtab!!!	neg	0,neu
55	Seda on nii halb lugeda. Ise püüan küll kõigi patsientide jutu mõistlikkuse piires ära kuulata neile otsa vaadata naeratada oma otsuseid seletada ja põhjendada aga alati see ilmselt ei õnnestu. Päris alati pole siiski patsiendil õigus! Kui keegi mulle samasugust loengut püüab pidada sotsiaalmaksust jne mille ma pean katkestama et vastuvõtuga edasi minna siis järgmise patsiendi jaoks olen vähemalt algul kindlasti kergelt ärritunud ja tõrges. Mida saab arst teha kui meditsiinisüsteem on meil sellisena paika pandud nagu ta on? Patsiendina eelistan meditsiinilisi eraettevõtteid kus minu kogemuse järgi on suhtumine patsienti tunduvalt inimlikum. Just sellest inimlikkusest jääb riigisüsteemis puudu.	pos	0,neu
56	http:renardmotorcycles.comindex01.phpKodukal on asi väga digitaalne...	neu	1
57	Jaa kõigil on ilmselt suured võlad Kuusmaa Saarjne. keskparteil on vene rahad taga muudkui ostab hingi ülesse.	neu	1
58	Vaene mees veab salakaupa.Rikas sama asjaga teeb äri.	neg	0,neu
59	kas korterid said otsa?	neu	1
60	..... Hää oli Win98SE.....nendele kes enam ei mäleta: win98 Õudus Kuubis Win98 ME More ErrorsWin98 SE Shit Edition95 2000 XP ja Win7 suht ok asjad.	neg	0,neu
61	http:www.postcrossing.com	neu	1
62	Ahv lippab metsas ringi ja kisab:Kriis! Kriis! Vastu tuleb hunt ja ütleb: Mis kriis ma sõin enne liha ja söön nüüd liha? Ahv jookseb kisades edasi: Kriis! Kriis! Vastu tuleb rebane või orav...? ja küsib: Mis kriis? Ma kandsin enne kasukat ja kannan ka edaspidi kasukat. Ahv jääb mõtlema:Tõesti mis kriis? Jooksin enne palja persega ringi ja jooksen ka edasi palja peega ringi.	neg	0,neu
63	Jürgen Veber võiks kah kätt proovida.	neu	1
64	Jääb mulje et pealkirjas NB! patuta naine ja kirjutise kõige viimases 25 realises lõigus ei hakka üle tsiteerima avaldub selgelt ja ühemõtteliselt kirjutise autori interpretatsioon ja hinnanguline seisukoht eetris olnud caseist ning rõhutatult neutraalne ei paista seeneed olevat.Aga pealkirja ja puändi vaheline tekst oli tore vahepala ikka.	pos	1
65	värsked puuviljad valmivad näe keegi on augu sisse tulistanud.	neu	1

66	ma ei tea kas Sul on endal lapsi või mitte. minul on. Ja minu vanem tütar on hetkel 14. Asi ei ole selles milline 14 aastane välja näeb sest kes meist ei tahaks selles vanuses juba täiskasvanulikum välja näha. Asi on selles et füüsiliselt sotsiaalselt ning vaimselt on 14 aastane siiski veel laps. Enamustes arenenud riikides ei ole lubatud TÄISKASVANU suguline vahekord nooremaga kui 18	neg	0,neu
67	Ma arvan et see on väga hea idee. Sooviksin kah näha kes selliste isiklikuks minevate kommentaaride taga tegelikult on :Edu Kätlyn! Kõik ongi nii nagu Sa kirjutasid.	pos	1
68	Mida öelda usumehele kes küsib kas su hing vajab Jumala armastuse puudutust?Kas valetada? Või sääred teha?kui su hing on tõesti nii õrnsiis tule usumehele otse näkku öelda:HOIA OMA RÄPASED KÄED MU ÕRNAST HINGEST EEMALE!	neu	0,neg
69	Kas puuke ei saa hävitada nagu muid kahjurputukaid näiteks nagu kassil kirpe tõrjutakse? Kas ei ole leiutatud inimesele kahjutut kuid puukidele surmavat mürki mida võiks pihustada riietele ja jalgadele? Kas puukide hävitamisega saaks looduslik tasakaal rohkem häiritud kui inimeste borrelioosi haigestumisega? Kas keegi oskab vastata?	neu	1
70	Sa mõtled ikka inglismaad!?!Sest sel korral noris muhku vaieldamatult briti lõvi!	neu	1
71	Jaa Perez oleks päris hea valik Massa asemele.	pos	1
72	jummel t6esti see jutt ajab ikka naermaalguses arvasin et see 1 aprilli nali	neu	1
73	Suurt reformi saab läbi viia minister kellel on usaldusmandaat.On ju täiesti mõttetu inimtööjõu ja raharaisk planeerida kokkuleppeid mingite tegelastega kelle ideoloogia on valetamisdoktriin.	neg	1
74	Jalavõru on minu arvates täiesti reaalselt teostatav võimalus.	neg	0,neu
75	F1 peaks ikka tipp tehnoloogia olema mitte mingi 60mili suurune liivakast	neg	1
76	ongi ju jama pank... üritasid üle oma varju hüpata aga ei õnnestunud. Pangatöötajad ei tee ühtegi liigutust ilma välismaalt juhatust saamata!!!	neg	1
77	Mannike erinevalt sinust ei ole ma hullunud parteifänn	neu	1
78	Miks küll EVs pääseb viimne inimsõnnik peaaegu alati ennetähtaegselt vabadusse??	neu	0,neg
79	evolutsioonhttp:www.listicles.comwpcontentupload...	neu	1
80	Kus on keerles ?	neu	1
81	Tegelikult on ju nii et kui avalik õigus süsteem ehk seadused ei suuda pervosid piisavalt kaua trllide taga hoida siis ei jäägi ju muud üle kui ohvri omastel asi ise ära teha vähemalt jääb selle inimese järgmine ohver olematta.	neu	1

82	Keegi ei saa iial kindel olla et meil selliseid värdjaid pole. Sellepärast peaks ajakirjandus olema eriti ettevaatlik selliste asjade avaldamisega. See võib anda mõnele meie hullule haiglase idee millegi samasugusega välja tulla. meie ajakirjanikud peaksid natukenegi oma peakesi tööle panema.	pos	0,neg
83	ah kuule pista oma piibel heaga sinna kohta! äkki hakkame kirikus käima ja palvetama..saame häkki kriisist välja ja 5 rikkama riigi sekka?	neu	0,neg
84	No niiiiiiiiiiiiii.Nüüd on eestlaste mark maailma turul täis situtud ja tänu isakese SAVISAARELE.Paras tallesee nägu peaks üldse keskerakonna areenilt ära kadumaja tema asemele peab tulema meeskes ütles legendaarsed sõnad:ME VÕIDAME NII KUI NII.Hr.Heinz Valk	pos	0,neg
85	to Lugeja 21.01.2009 06:29mis oli artikli mõte?Kuskil hõõgub hüsteerialõke ja keegi peab ju seda õhutama...	neu	1
86	vahel päris huvitav eriti Vaba MÕtte Klubi viimase sõnaga võin eksida. Kui sed avaatan siis ei kujuta küll ette mis juhtuks kui üks selline saade käiks ka läbi ETV. Siis vis lastaske kõik ETV juhid paugupelat kohalt lahti	pos	0,neu
87	Õpetajatest on ikka väga palju vaesemaid!!!	neu	1
88	Artiklis oli juttu sellest et kesikud lasid põhjavee toiteala ürgoru täis ehitada mis muudab sealt põhjaveevõtu peaaegu võimatuks reostunud. Muus maailmas üldiselt kaitstakse põhjavee toitealasid meil aga ruulib kinnisvaraarendus ja majanduskasv. Seda siis ka Männiku liiviku puhul mida varem peeti Tallinna alternatiivseks veevarustusallikaks kui Ülemistes midagi juhtuks. Nüüd on aga kaevanduslubadega seegi hävimisele määratud.Tln Vee ärastamine ei puutugi niipalju asjasse v.a see et TV omanikele on odavam joogivett pinnaveest toota kui põhjaveest. Ja seetõttu me kõik seda kloorisegu luristamegi.	pos	0,neg
89	kommunistide tapmine nii koonduslaagris kui ka mujal on alati igati õigustatud! Kui ei tapa kommunisti siis tapab kindlasti tema ise teid!!!	neg	1
90	No on nad seal enne midagi asjalikku teinud? Praegu vist Rimis pole kampaaniat et kõik pereemad kleepse kleebiks mis vaesekestel siis muud üle jääb kui delfis istuda ja omale vabandavaidõigustavaidülistavaid komme trükkida kes see teine seda nende eest teeb vat see on nende raske töö	neg	1
91	Ma arvasin esialgu et rattaga ületas kiirust. Autoga on köömes iga könn teeb järgi :D	neu	1
92	Teate kui mina oleks mees ja oleks arisoont ja korralikult pappi oleks ma votaks ka sellise naise kes ytleb et jahid kyll meeldivad aga kui raha pole siis ujun ka monuga. Miks on vaja naist halvustada? Kas rikastele peaks armastus olema keelatud? Ysna inetu on vordsustada naisterahva ja ema elu vaartust elu loteriiga.	neu	1

93	kui tegemist on omadussõnaga siis peaks ikka ütlema murettekitav	neu	1
94	Mina suhtun hinnatõusu positiivselt. Tallnna linnavalitsus on valitud demokraatlikul teel ja väljendab valijate tahet. Keskerakond hoolib.	pos	1
95	eestlastel oli vast ikka näärid mis tähendasid valguse võitu pimeduse üle?	neu	1
96	Saksamaal tehtud uuringutest selgub et vaid 10 pedofiilidest kordavad oma kuritegu.Statistika teatavasti ei valeta.http:www.rahvaraamat.eeindex.php?id62...	neu	1
97	Ma tõesti ei saa sellest artiklist midagi aru. Tundub nagu mõni viie aastane on seda teil seal tõlkinud:	neu	1
98	Võimalik et maiad lihtsalt ei viitsinud järgmist päikeseaastat üles kribada.	neu	1
99	isegi kui ka kommentaarid on asjalikud on selline artikkel ka juba ammu oodatud. kui artikkel kutsubki üles debatile siis tehkem seda. veel rohkem tuleb asjast rääkida! K.Paulus imestab miks sots.reklaami maha tehakse...tavaline inimene tänavalt ei mõtlegi ju samamoodi nagu reklaamiproff aga kui asju selgitada siis suhtumine muutub nagu ka haigete kohta annetuste kohta jne ka artiklis öeldud. aitähh K. Paulus. mina hakkasin sotsiaalreklaamile leebemalt vaatama kui seni.	neu	1
100	Politsei poolt olid ohutustehnika reeglid täitmata.Vahet pole kes süüdi.Oleks reegleid jälgitud oleks kõik elus.Oleks tädil rattad jne.	neu	1

Lisa 2: rakendus

Rakendus on kättesaadav CD'l ja lõputööde registrist. CD on lisatud paberdokumendi lõppu.