TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond
Informaatikainstituut
Tarkvaratehnika õppetool

Meetod uudiste voogude tonaalsuse hindamiseks

Magistritöö

Üliõpilane: Kaur Laanemäe
Üliõpilaskood: 132306IABMM
Juhendaja: Ahti Lohk

Tallinn
2015

Autorideklaratsioon

Kinnitan, et olen koostanud antud lõputöö iseseisvalt ning seda ei ole kellegi teise poolt varem kaitsmisele esitatud. Kõik töö koostamisel kasutatud teiste autorite tööd, olulised seisukohad, kirjandusallikatest ja mujalt pärinevad andmed on töös viidatud.

(*kuupäev*) (*allkiri*)

Annotatsioon

Antud magistritöö eesmärgiks on luua meetod reaalajaliste uudiste analüüsiks valitud uudiste portaalide tekstidel koos ettevõtete nimede ja märksõnade eraldamisega iga uudise kohta.

Töös pakutakse välja uus tonaalsuse hindamise meetod, uuritakse teksti pealkirjast ettevõttenime eraldamise võimalusi ja tutvustatakse vahendit märksõnade leidmiseks tekstist.

Tonaalsuse leidmisel kombineeritakse kahte erinevat leksikaalset ressurssi ja tõdetakse, et tulemus on märgatavalt täpsem kui neid ressursse eraldi rakendades. Ettevõtete nime leidmisel kasutatakse spetsiaalset keeletöötlusvahendit Standford NER ja töö autori poolt testiks loodud börsiettevõtete nimekirja. Märksõnade eraldamisel rakendatakse aga mitmeotstarbelist ja automaatset teemade indekseerijat ehk Maui-indexeri-nimelist keeletöötlusvahendit.

Töös väljapakutud meetod on hiljem aluseks plaanitava idufirma teenuse väljatöötamisel.

Lõputöö on kirjutatud eesti keeles ning sisaldab teksti 54 leheküljel, 7 peatükki, 10 joonist ja 10 tabelit.

Abstract

The aim of this thesis is to create a method for real-time news analysis from selected news portals with company names and keywords from each news text.

Central problem of this thesis is finding polarity of news. For which there was proposed a new method for finding polarity of the news. This thesis also examines ways to find organization name from the news title and presents a tool for finding keywords from the text.

Method to find the polarity of the news text combined two different lexical resources where it is found that the outcome is much more accurate than using these resources separately. To find the company names a language processing tool named Standford NER and a list of companies created by the author are used. For finding the keywords a multi-purpose automatic topic indexing tool, in other words maui-indexer tool is used.

Proposed method will be later used in the development of a news analysis service.

The thesis is in Estonian and contains 54 pages of text, 7 chapters, 10 figures and 10 tables.

Lühendite ja mõistete sõnastik

NLP	***Natural language processing***
	Loomuliku keele töötlus
POS Tagger	***Part-Of-Speech Tagger***
	Sõnaliigi märgendaja
NER	***Named Entity Recognizer***
	Nime üksuste tuvastaja

Jooniste nimekiri

Joonis 1 - Teksti tonaalsuse leidmise esialgne protsess 20
Joonis 2 - Teksti analüüsiks kulunud aeg 21
Joonis 3 - Ettevõtte nime leidmine 26
Joonis 4 - Enimkasutatud sõnade leidmise protsess 28
Joonis 5 - Analüüsi tulemused (sõnade keskmine) 30
Joonis 6 - Analüüsi tulemused (sõnade summa) 31
Joonis 7 - Analüüsi tulemused (kahe meetodi keskmine) 33
Joonis 8 - Esimene tonaalsuste jagamise meetod 34
Joonis 9 - Märksõnade tonaalsus (sõnade summa) 37
Joonis 10 - Teksti ja märksõnade arvutatud keskmised 38

Tabelite nimekiri

Tabel 1 - Loomuliku keele töötluse raamistike võrdlus. 16
Tabel 2 - Stanford Named Entity Tagger tulemused 25
Tabel 3 - Meetodite vea protsent 35
Tabel 4 - Vastuoluliste uudiste märksõnade tonaalsus 36
Tabel 5 - Enimkasutatud sõnade tonaalsuste summa 39
Tabel 6 - Stanford Log-linear Part-Of-Speech Tagger märgendid 45
Tabel 7 - Positiivsete uudiste märksõnad 46
Tabel 8 - Negatiivsete uudiste märksõnad 47
Tabel 9 - Kauguse arvutamise tulemused 51
Tabel 10 - Tonaalsuse analüüsi väljundid 53

Sisukord

1. Sissejuhatus 10
1.1 Taust ja probleem 10
1.2 Ülesande püstitus 11
1.3 Hetkeseis maailmas 11
1.4 Metoodika 12
1.5 Ülevaade tööst 13
2. Kasutatavad tehnoloogiad 14
2.1 Sõnade andmebaasid 14
2.2 Loomuliku keele töötlus 15
2.3 Märksõnade leidmine tekstist 17
3. Teksti tonaalsuse leidmine 18
3.1 Esmane sisu tonaalsuse hindamine ja väljundid 18
3.2 Programmi jõudlus 21
4. Mittekvantitatiivsed näitajad 23
4.1 Kuidas leida ettevõtte nimi tekstist 24
4.2 Märksõnad tekstist 27
4.3 Enimkasutatud sõnad tekstis 27
5. Tulemuste hindamine 29
5.1 Meetodi valik 29
5.2 Tonaalsuse leidmine 34
5.3 Tulemuste võrdlus 35
6. Mittekvantitatiivsete näitajate seos tonaalsusega 36
6.1 Märksõnad ja tonaalsus 36
6.2 Enimkasutatud sõnad ja tonaalsus 39
7. Kokkuvõte 40
Summary 42
Kasutatud kirjandus 43
Lisa 1 45
Lisa 2 46
Lisa 3 47

Lisa 4 48
Lisa 5 51
Lisa 6 53

1. Sissejuhatus

Käesoleva töö eesmärk on leida kiire viis suure hulga kirjalikus meedias avaldatud uudiste tonaalsuse analüüsiks ning välja selgitada, missuguse ettevõtte kohta konkreetne uudis käib. Peale selle tuleb pakkuda lähenemine märksõnade eraldamiseks. Selle eesmärk on anda inimestele kiire ülevaade uudise või uudiste kohta, mis käsitlevad mingit kindlat ettevõtet või ettevõtlussektorit käesoleval ajahetkel.

Antud töö on tehtud 2015. aasta esimesel poolel ning selle alusel on kavas luua baas uuele idufirmale, mille tegevusala on uudiste reaalajas analüüs ning selle info müümine teenusena.

1.1 Taust ja probleem

Mitmed autorid on märkinud, et ettevõtete kohta käivad uudised mõjutavad aktsiate hinda väärtpaberiturul [1]. Sealjuures halvad uudised mõjutavad aktsia hinda rohkem kui head uudised [2]. Teada on veel, et kõige suuremad aktsiahinna muutused toimuvad 20 minutit enne ja pärast esimeste uudiste ilmumist [3]. Pärast esialgset suurt hüpet aktsiahinnas lühikese aja jooksul järgneb pikemaajaline liikumine esialgsele hüppele vastupidises suunas [4].

Eelmainitud asjaolu arvestades eeldame, et uudise tonaalsuse analüüs uudise ilmumise hetkel aitab kaasa kiirete otsuste vastuvõtmisel väärtpaberiturul. Autori parima teadmise juures sellist teenust hetkel ei eksisteeri.

Eeltoodud teenuse potentsiaalseteks klientideks võiks olla kõik väärtpaberitega kauplevad isikud, keda on ainuüksi Ameerika Ühendriikide turul eeldatavasti 7,5 miljonit. Antud teenusel on veel teisi potentsiaalseid kasutusvõimalusi nagu näiteks finantsuudiste kiire ülevaate andmine ajakirjanikele, kus loeb see, kui kiiresti uudised nende uudiste portaali jõuavad. Seega on tegemist potentsiaalselt tootega, millist võiks üle maailma vajada kümned miljonid inimesed.

1.2 Ülesande püstitus

Töö kõige olulisemas eemärgiks on leida meetod uudiste sisu analüüsiks, mis tagastaks lihtsa numbrilise tulemuse uudise sisu hoiaku kohta. Antud juhul huvitab meid, kas uudis on vaadeldava ettevõtte jaoks positiivse või negatiivse sisuga.

Meetodi headuse testimiseks tuleb saadud tulemusi võrrelda inimeste antud hinnangutega ja välja selgitada meetodi täpsus. Peale selle tuleb loodud meetodit võrrelda teiste autorite loodud lähenemisviisidega. Siinjuures on eesmärgiks saada teistest meetoditest parem tulemus, mille hinnangu saamiseks tuleb võrrelda töö autori ja teiste meetodite tulemuste vigade arvu.

Lisaks tuleb leida iga uudise kohta selle uudisega seonduvad märksõnad. Ka selle osa puhul tuleb vaadata, kas märksõnade ja tekstist leitud inimese arvamuse vahel on seos, kui kasutada uudiste sisu analüüsi asemel ainult märksõnu. Lisaks on märksõnad vajalikud inimesele kiireks info kuvamiseks.

Viimasena on meil vaja teada, missuguse ettevõtte kohta avaldatud uudis käib. Seda on vaja nii andmete koondamiseks kui ka edasiseks analüüsiks, mida selles töös ei vaadelda.

1.3 Hetkeseis maailmas

Hoiakute kaevandamine on tänasel päeval järjest rohkem uuritav teema. Peamisteks lähenemisteks on leksikaalsete sõnakogumite kasutamine [5, 6, 7, 8] või masinõppe kasutamine [7, 9, 10].

Leksikaalsed ressursid, mille kasutamist on uuritud, on SentiWordNet [6, 7, 8], OpinionFinder Lexicon [8] ja AFINN [7, 8].

On olemas juba töötavaid teksti tonaalsuse analüüsi meetodeid ja ressursse. Vaadeldud on selliseid süsteeme nagu SentiStrength [8] ja Sentiment140 [8]. Mõlemad neist on mõeldud lühikeste mitteformaalsete tekstide nagu *twitter*'i säutsude tonaalsuse analüüsiks.

Samuti on kombineeritud erinevaid leksikaalseid ressursse *twitter*'i postituste tonaalsuse analüüsiks, mille tulemusena on saadud otsustuspuu, mille järgi saab teada, mis järjekorras vaadeldud ressursse kasutada, et saada täpseim tulemus [8].

Masinõppe lahenduse juures on kindlasti üheks enamlevinud lahenduseks naive Bayes'i klassifikaatori kasutamine [7, 9, 10]. Lisaks on masinõpet kasutatud erinevate ressursside kombineerimisel, et leida parim viis nende järjestamiseks ja kasutamiseks [8].

Kuigi tegemist on teemaga, mida on uuritud kasutades mitmeid erinevaid lähenemisi, pole olemas ühtegi meetodit, mille abil oleks võimalik määrata tulemus 100%-lise täpsusega. Paistab, et kuigi meetodeid on kombineeritud, pole leksikaalseid ressursse kombineeritud viisil, kus mitme ressursi tulemust ühe tulemuse arvutamiseks korraga kasutatakse.

1.4 Metoodika

Antud toos kasutatakse uudiste sisu analüüsiks kahte erinevat leksikaalset ressurssi. Nendeks on SentiWordNet ja AFINN, mille abil leitakse kaks erinevat kvantitatiivset tulemust ühe teksti kohta.

Uudise abil leitakse erinevaid mittekvantitatiivseid andmeid, millest on tulevikus loodava teenuse jaoks kasu. Uudise pealkirjas olevate ettevõtte nimede tuvastamiseks kasutatakse Stanford NER algoritmi ja ettevõtete nimede andmebaasi. Lisaks leitakse maui-indexer'i raamistiku abil tekstist märksõnad. Samuti pakutakse välja lihtne võimalus enimkasutatud sõnade leidmiseks tekstist.

Teksti sisust saadud tulemusi võrreldakse inimese antud teksti tonaalsustega ning nende teadmiste abil leitakse lihtne meetod, kuidas leida üks tulemus kasutades kahte erinevat sõnakogumit. Lihtsustatult on meetodiks graafikul leitud joon, mis võiks kujutada endast teksti, mis ei ole positiivse ega negatiivse tonaalsusega. Selle abil arvutati saadud punkti kaugus loodud null-joonest, mis on ühtlasi uus kombineeritud teksti tonaalsus.

Veendumaks, et loodud meetod toimib paremini, kui kummagi leksikaalse ressurssi eraldi kasutamine, võrreldakse erinevatel juhtudel nii SentiWordNet kui ka AFINN abil saadud tulemuste vigade arvu loodud kombineeritud meetodiga.

1.5 Ülevaade tööst

Antud alampeatükis anname töö lühiülevaate peatükkide kaupa.

Peatükk 2 pakub potentsiaalsed tehnoloogiad kolme erineva ülesande lahendamiseks - tonaalsuse tuvastamiseks, ettevõtete nimede ja märksõnade eraldamiseks tekstist.

Peatükk 3 käsitleb, kuidas kahte erinevat leksikaalset resurssi eraldi kasutada ja nende abil saada kummalgi juhul tulemused. Lisaks uurime, kas loodud meetod töötab piisava kiirusega, et seda oleks tulevikus mõistlik kasutada.

Peatükk 4 on pühendatud mittekvatnitatiivsete näitajate leidmisele. Nendeks on ettevõtte nimi, mille kohta uudis käib, märksõnad tekstist ning enimkasutatud sõnade nimekiri.

Peatükk 5 uurib erinevaid meetodeid, kuidas saab kasutada kolmandas peatükis „Teksti tonaalsuse leidmine“, saadud tulemusi. Leitakse parim meetod, kuidas kahe erineva meetodiga saadud tulemusi võrrelda ning nende abil leida üks tulemus, mis näitab ära teksti tonaalsuse. Saadud tulemust võrreldakse kummagi kasutatud meetodiga eraldi, kus leitakse, et meetod, mis kombineerib tulemusi, toimib paremini kui kumbki meetod eraldi.

Peatükk 6 on pühendatud mittekvantitatiivsete näitajate ja teksti tonaalsuse vaheliste seoste uurimisele.

Peatükk 7 võtab kokku töö olulisimad tulemused ja kirjeldab pakutud meetodi edasiarendamise võimalusi.

2. Kasutatavad tehnoloogiad

On teada, missuguseid ülesanded lahendust vajavad ning ka mida on eelnevalt sarnaste probleemide lahendamiseks tehtud. Sellest tulenevalt uurime välja, milliseid tehnoloogiaid ja ressursse me töös kasutada võiksime, et saavutada parem tulemus, kui minevikus tehtud tööde tulemus oli.

Selle töö raames on tarvis leida tehnoloogiad kolme erineva ülesande lahenduseks. Esimene ja ühtlasi kõige tähtsam osa on teksti tonaalsuse leidmine. Teiseks on vaja leida loomuliku keele töötluse raamistik, mida on arvatavasti vaja teksti polaarsuse leidmise juures kui ka pealkirjast ettevõtete või organisatsioonide nimede leidmise jaoks. Viimaseks osaks on tekstist märksõnade leidmine, mille jaoks on vaja eraldi lahendust.

2.1 Sõnade andmebaasid

Kasutame töös kahte erinevat leksikaalset ressurssi ning proovime leida viisi, kuidas neid kombineerida, et tulemus oleks täpsem kui kummagi ressursi abil saadud tulemus eraldi. Eeldame, et sellisel juhul on võimalik leida täpsem tulemus, kuna erinevad ressursid on saadud erinevate meetodite abil ning sel juhul on võimalik kombineeritud tulemus tasakaalukam ja seega ka täpsem. Lisaks suurendab selline meetod ka sõna leidmise tõenäosust, kuna kahe ressursi puhul on kombineeritud sõnade arv suurem, sest need ei pruugi täielikult kattuda. Ka see võib aidata kaasa täpsuse suurenemisele.

Lisaks võib kahe meetodi puhul arvata, et kui tulemused mõlemal juhul näitavad sarnast teksti tonaalsust, siis on ka inimese arvamus suurema tõenäosusega sama tonaalsusega, mis mõlema meetodi tulemi puhul. Erinevate tulemuste korral võib eeldada, et kombineeritud tulemus on aga väiksema tõenäosusega vigane.

Esimeseks ressursiks on SentiWordNet v3.0.0'i [11], mis on WordNet'i [12] laiendus. Tegemist on leksikaalse andmebaasiga, mis on loodud arvamuse kaevandamiseks. Antud andmebaas on jaotatud CC BY-SA 3.0 litsentsi [13] alusel. Meid huvitab olemasolevatest andmetest sõnatüüp, positiivsuse ja negatiivsuse skoor ning sõna ise. Sõnatüüpe on antud andmebaasis kasutusel neli: nimisõna, omadussõna, määrsõna ning tegusõna. Peale sõna

tonaalsuse on lisaks antud ressursis olemas ka sõna objektiivsuse skoor. Selles töös uudise objektiivust ei hinnata, seega seda osa ressursist ei kasutata.

Teisel juhul kasutame AFINN [14] sõnade nimekirja. Tegemist on aastatel 2009-2011 Finn Årup Nielsen poolt loodud sõnade nimekirjaga. Igale sõnale on antud väärtus miinus viie ja pluss viie vahel, kus negatiivne väärtus tähendab negatiivset sõna tonaalsust ja positiivsete väärtuste korral on tegemist sõnaga, mis on ühtlasi positiivse tonaalsusega. Igale sõnale on väärtus antud käsitsi ning see ressurss keskendub sõnadele, mida kasutatakse mikroblogides nagu näiteks *twitter*. Lisaks sisaldab AFINN ka slängisõnu, mida ei pruugi olla sõnaraamatutes. Sõnade nimekiri on jaotatud ODbL v1.0 litsentsi [15] alusel. Täpsemalt kasutame kõige viimast versiooni nimega AFINN-111, kus on 2477 sõna.

Uudiste kommentaaride puhul on osutunud kõige täpsemaks meetodiks AFINN sõnade nimekirja kasutamine. [7] Samas ei pruugi see nii olla tervete uudiste analüüsi puhul, kuna kommentaarides on tavaliselt kasutatud vabamat sõnavara, mida leiab suurema tõenäosusega just sellest ressursist. Seega ei hakka me kumbagi meetodit välistama ning antud juhul võib eeldada, et need ressursid just täiendavad üksteist.

2.2 Loomuliku keele töötlus

SentiWordNet v3.0.0'i kasutamiseks on vaja teada sõna tüüpi ning lisaks ka sõna algvormi. Selle jaoks on vaja leida mõni loomuliku keele töötluse raamistik. Lisaks on vaja leida pealkirjadest ettevõtete nimed, kus võiks ka samast raamistikust kasu olla.

Esimene raamistik, mida vaatleme, on Stanford NLP [16]. Tegemist on Stanfordi Ülikoolis arendatava vabavaralise raamistikuga mis on litsentsitud GPL versioon 3 [17] alusel. Antud raamistik paistab esialgu toetavat kõiki vajaminevaid funktsioone.

Teine raamistik, mille puhul on näha, et see toetab kõiki funktsioone, on Apache OpenNLP [18]. OpenNLP puhul on tegemist vabavaralise raamistikuga, mis on litsentsitud Apache litsentsi versioon 2.0 [19] alusel.

Kolmas raamistik, mida vaatame, on OpeNER projekt [20], mis on rahastatud Euroopa Komisjoni poolt ning on litsentsitud Apache litsentsi versioon 2.0 [19] alusel.

Eelnevaid raamistikke võrreldes vaatleme potentsiaalselt vajaminevate funktsioonide olemasolu kui ka kasutatavaid tehnoloogiaid.

Tabel 1 - Loomuliku keele töötluse raamistike võrdlus.

	StanfordNLP	**OpenNLP**	**OpeNER**
POS Tagger	Jah	Jah	Jah
NER	Jah	Jah	Jah
Tokenization	Jah	Jah	Jah
Lemmatisation (Sõna algvormi ledimine)	Jah	**Ei**	Jah
Lauseks jagamine	Jah	Jah	Jah
Kasutatud tehnoloogiad:	Java	Java	Ruby, Python, Java, Perl

Valikus olevatest loomuliku keele töötluse raamistikest valime edaspidise töö jaoks StanfordNLP, kuna see toetab kõiki esmapilgul vajaminevaid funktsioone ning kasutab keeleliselt ühte kindlalt programmeerimise keelt. OpenNLP puhul ei ole teada, et see toetaks sõna algvormi leidmist. OpeNER probleemiks on aga, et seal kasutatakse nelja erinevat programmeerimise keelt ning programmi tööks peab ka süsteemis olema nende kõigi kasutamise tugi.

Stanford NLP raamistik pakub kasutajale statistilisi NLP, süvaõppe NLP kui ka reeglite põhiseid NLP tööriistu erinevate lingvistiliste probleemide lahendamiseks. Kõige rohkem huvitab meid kolm osa POS - märgendamine ehk sõnaliigi leidmine, NER märgendamine ehk nimeliste objektide leidmine kui ka sõna algvormi leidmine.

Sõnaliigi leidmiseks on antud raamistikus loodud Java implementatsioon kahest tööst [21, 22] samal teemal.

Nimeliste objektide tekstist leidmine funktsionaalsus baseerub kaudselt Jenny Rose Finkel, Trond Grenager ja Christopher Manningu artiklil [23].

2.3 Märksõnade leidmine tekstist

Tarkvara kasutajale on vaja tekstist parema ülevaate andmiseks leida märksõnad. Ükski loomuliku keele töötluse raamistik tekstist märksõnu eraldi leida ei oska. Seega on vaja ka selle ülesande jaoks leida mõni raamistik.

Esimene raamistik, mida vaataleme, on Kea. [24] Tegemist on algoritmiga, mille eesmärk on leida tekstis märksõnu ja fraase. Antud programm on kirjutatud Java's ning on litsentsitud GPL versioon 3 [17] alusel.

Teine raamistik on maui-indexer [25]. Ka see on litsentsitud GPL versioon 3 [17] alusel. Raamistik ise on kirjutatud programmeerimise keeles Java ning kui vaadata selle osi, siis eelnevalt vaadeldud algoritm on selle raamistiku üks osa. [26]

Kuna üks leitud algoritm on teise raamistiku üheks osaks, siis valime selleks juhuks teise leitud raamistiku.

Maui-indexer'i raamistiku märksõnu võrreldes 332-e inimesega on kokkulangevus 0% kuni 80%, kus keskmiseks kokkulangevuse määraks on 23,8% [27] ning ainukesed juhud, kus oli väga väike kokkulangevus, olid juhud, kui inimene oli märkinud vähe märksõnu (üks kuni kolm). Seega võib öelda, et tegemist on piisavalt täpse viisiga märksõnade leidmiseks loodava tarkvara jaoks. Seda sellepärast, et märksõnad on lisainfoks inimestele ning neil on võimalik soovi korral ka lugeda terviklikke artikleid.

3. Teksti tonaalsuse leidmine

Teksti tonaalsuse leidmine on käesoleva töö tähtsaim probleem. Kuna selle tulemuseks on mingi hulk numbrilisi näitajaid teksti kohta, siis on nende abil võimalik analüüsida teksti ning teha järeldusi. Samas peab teksti analüüsimine olema kiire tegevus, kuna uudiste voog maailmas on suhteliselt suur ning päeva tipphetkedel võib uudiseid tulla korraga väga palju. Seega ei saa olla teksti polaarsuse hindamise protsess väga keerukas.

Lühikese aja all mõtleme protsessi, mille täitmiseks ei kulu keskmiselt üle sekundi. See on piisavalt lühike aeg, et serveris ei tekiks ka päeva tipphetkedel viivitusi, mida inimesed võiksid tähele panema hakata. Seega seame üheks nõudeks, et uudis oleks enne analüüsitud, kui inimene seda vastavas uudiste portaalis lugema oleks jõudnud hakata või veelgi parem, saaks uue uudise olemasolust teada loodava teenuse vahendusel.

Nendest nõuetest tulenevalt pakume välja protsessi, mille alusel võiks analüüsida uudiseid piisavalt lühikese ajaga. Protsessi tulemus on numbriliste näitajate paar, kus üks number on saadud kasutades SentiWordNet'i ja teine AFINN sõnade kogumit. Sellist tulemuste paari on võimalik analüüsida edaspidises töös ning teha nendest järeldusi võrreldes inimese pakutud teksti tonaalsusega.

Kahte erinevat sõnade kogumit kasutame eeldusel, et antud leksikaalsed ressursid täiendavad teineteist, st suurendavad nii-öelda leksikaalset katet ja mis omakorda garanteerib tulemuse suurema täpsuse kui kasutada vaid ühte ressurssidest.

3.1 Esmane sisu tonaalsuse hindamine ja väljundid

Tonaalsuse hindamiseks kasutame Stanford NLP keeletöötluse raamistikku ning kahte erinevat sõnade kogumit, milleks on SentiWordNet 3.0.0 ja AFINN-111.

Esimese osana vaatleme SentiWordNet'i tulemuse leidmist. Et leida sõna antud sõnade kogumikust, on vaja leida teksti kontekstist info otsitava sõna kohta, mille jaoks on tarvis kasutada Stanford NLP raamistikku.

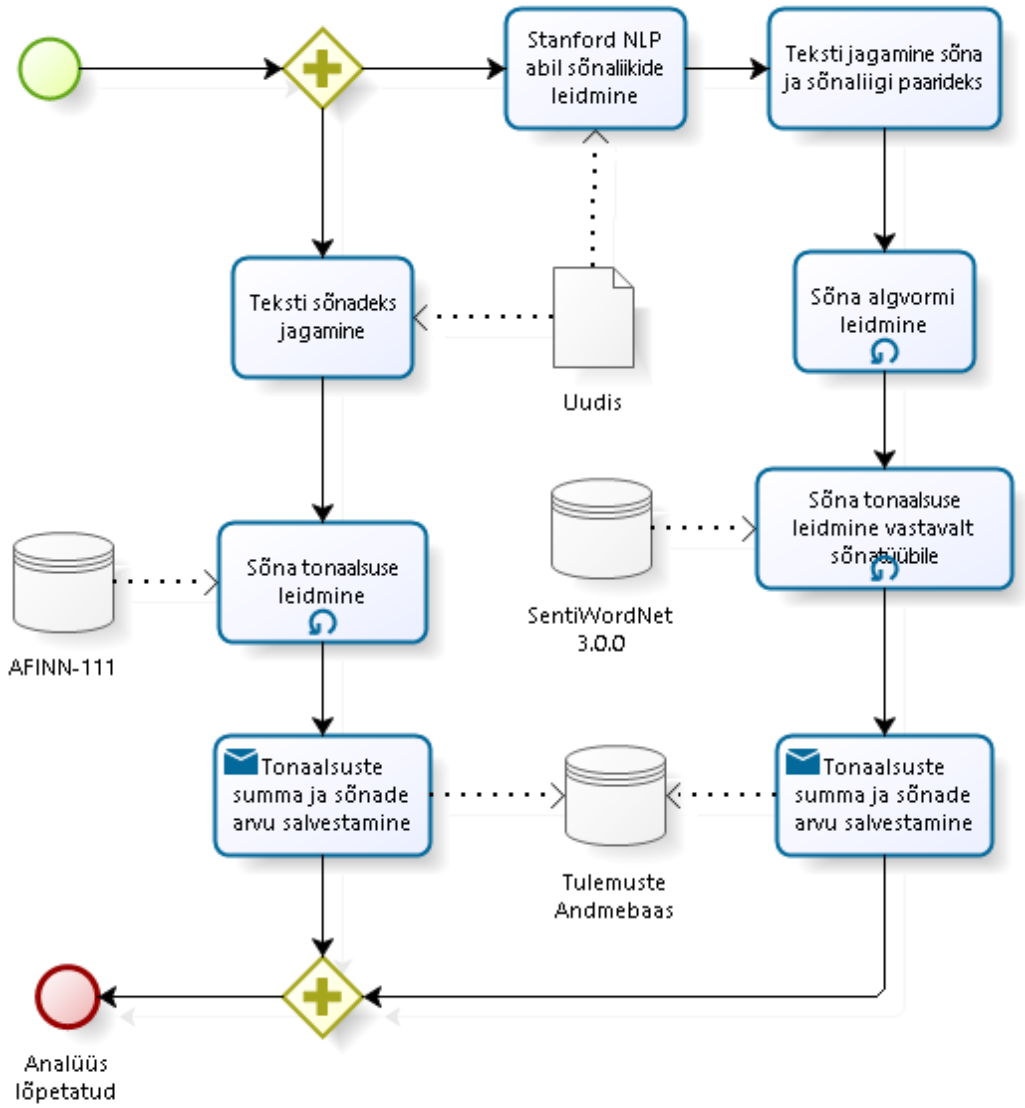
Kõigepealt on meil vaja teada sõna liiki. Tulemuseks on sõna ja temaga seotud märgend, mille variante on *Stanford NLP POS tagger*'il 36. (vt Lisa 2). *SentiWordNet*'il on neid aga neli. Seega tuleb leitud märgendid jagada nelja suuremasse gruppi.

Edasi on SentiWordNet'i kasutamiseks vaja teada ka sõna algvormi. Selle saamiseks kasutame samuti *Stanford NLP* raamistikku. See funktsionaalsus on antud keeletöötluse raamistiku tuumikfunktsionaalsuses olemas.

Seega, kui on teada nii sõna algvorm kui ka sõnatüüp, on meil võimalik kasutada *SentiWordNet*'i ning selle abil leida sõna tonaalsus.

Saadud sõnade tonaalsused liidame kokku ning jätame meelde, mitu sõna tekstis oli ning need salvestame edasiseks analüüsiks.

Järgmiseks kasutame AFINN-111 sõnade andmebaasi, mille sisendiks on töötlemata tekst. Antud juhul kasutame töötlemata teksti, kuna antud sõnade kogumis on sõnad juba sellises vormis, mida tekstist leida võib. Väljundiks jätame meelde sõnade summa ja ka sõnade arvu, mida analüüsiti.



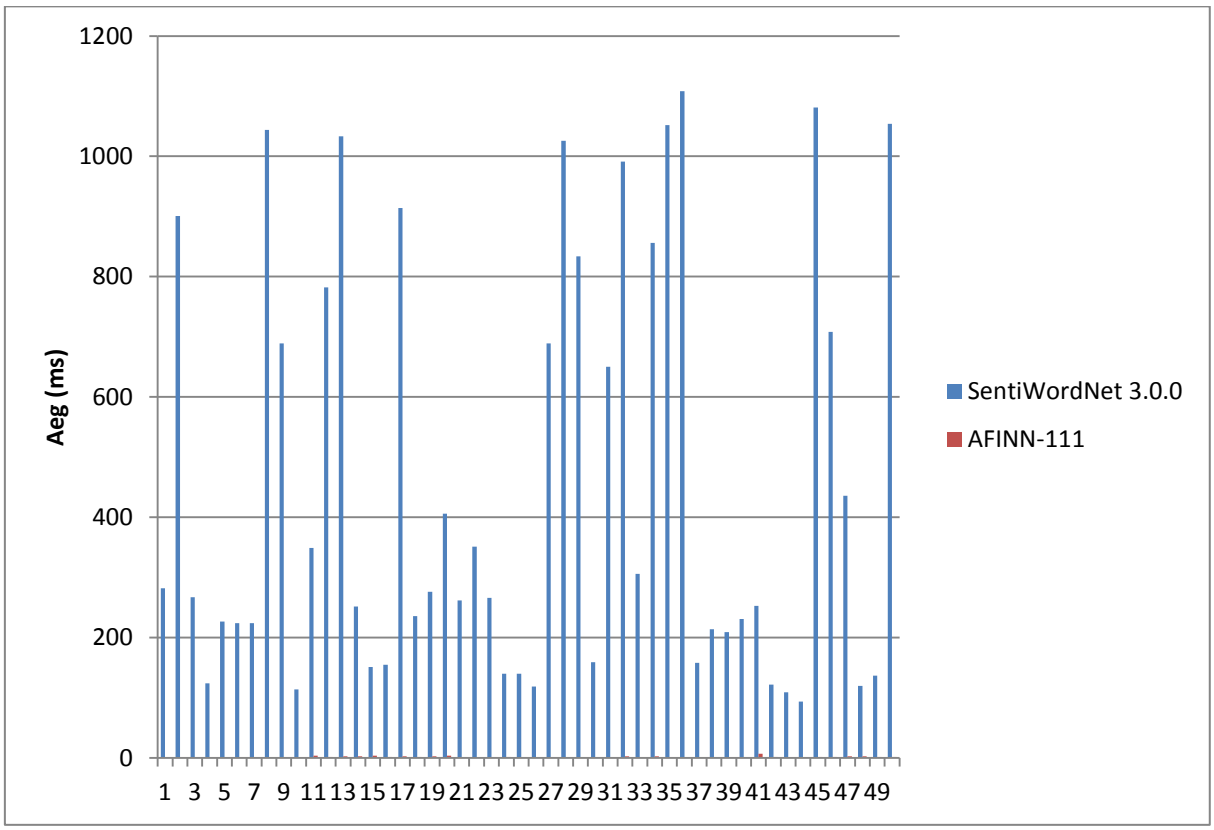
Joonis 1 - Teksti tonaalsuse leidmise esialgne protsess

Joonisel (Joonis 1 - Teksti tonaalsuse leidmise esialgne protsess) on näha esialgse uudiste analüüsi protsessi diagramm.

Antud protsessi kahe erineva tulemuse abil võiks olla võimalik hakata arvutama ühtset tulemust, mida inimesele kuvada võib. Selleks on vaja kõigepealt sooritada võrdlus inimese arvamusega tekstidest ning seejärel leida parim viis ühtse tulemuse arvutamiseks, mida võib kasutajale kuvada.

3.2 Programmi jõudlus

Programmi töös tasub kindlasti ka vaadata, kui kaua kulub aega ühe uudise analüüsiks. Selleks vaatleme viiekümne uudise analüüsiks kulunud aega. Programmi töös on tähtis, et uudise analüüs oleks piisavalt kiire tegevus, et serveris ei tekiks uudiseid analüüsides järjekorda.



Joonis 2 - Teksti analüüsiks kulunud aeg

Nagu graafikult näha on, kulub AFINN-111 listi kasutades aega kõige vähem. Seda põhjusel, et antud osa juures pole vajalik teksti töötlus, seega listist ühe sõna leidmise keerukus on O(log n). Kuna sõnu on tekstis rohkem kui üks, võib öelda, et kogu algoritmi keerukus antud juhul on O(n log n).

SentiWordNet 3.0.0 listi kasutamiseks on vajalik eelnev teksti töötlus. Nagu graafikult näha, kulub selle osa jaoks märgatavalt rohkem aega. Kõige rohkem kulus uudise analüüsiks 1108 millisekundit. Keskmine aeg oli 450,5 millisekundit ja mediaanajaks oli 266,5 millisekundit. Selline viivitus oleks juba inimese jaoks märgatav.

Samas tuleb arvestada, et kogu see tegevus toimub tausta poolel, seega inimese jaoks ei tähenda see midagi muud, kui et uudise analüüsi tulemused tulevad teenuses nähtavale sekund pärast uudise avalikuks tulemist.

Teise osana tuleb arvestada, et tegemist on ühe protsessori lõime kasutamisega. Kuna tänapäeva serveritel on minimaalselt 6 tuuma ja tavaliselt tuuma kohta 2 lõime, siis ka ühe lihtsama serveriga on võimalik analüüsida korraga vähemalt 12 uudist. Arvestades seda, et uudiseid ühes portaalis ei tule mitu tükki sekundis ning isegi näiteks viitkümmet maailma populaarsemat uudisteportaali korraga jälgides ei tohiks uudise väljastamise ning analüüsi tulemuste vahe olla rohkem kui mõned sekundid.

Järelikult tähendab see, et kui loodav teenus uuendaks kasutajale kuvatavaid tulemusi reaalajas vastavalt sellele, kui tulemused tausta poolelt valmis saavad, ei tohiks tekkida ühtegi juhtu, kus inimene suudab uudist enne lugeda, kui see loodava teenuse poolt juba analüüsitud on.

Pärast antud analüüsi andmete konsolideerimine on ülesanne, mille puhul ei teki viiteid juhul, kui andmebaasi ülesehituses on mõeldud erinevatele seostele, mille järgi infot otsitakse.

4. Mittekvantitatiivsed näitajad

Peale tekstist leitud tonaalsuse näitajate on loodava süsteemi jaoks vaja teada saada uudise kohta veel informatsiooni, mis ei oma numbrilisi näitajaid. Selle eesmärk oleks anda lisainfot isikutele, kes loodavat süsteemi kasutaksid, kuid tegemist oleks infoga, millest ei pruugi olla arvutuslikus analüüsis kasu. Selliseks infoks oleks kindlasti märksõnad ja võib-olla ka enimkasutatud sõnade nimekiri. Lisaks on vaja leida infot, millest pole küll tulemuste arvutamisel kasu, kuid mida on vaja programmi tulevases töös. Selliseks infoks on ettevõtte nimi, mille abil otsib inimene infot või teeb loodav teenus mingisuguseid kokkuvõttvaid analüüse.

Seega tuleb välja, et kõige tähtsamaks selliseks infoks, mida loodaval süsteemil vaja läheb, on ettevõtte nimi. Antud infot on vaja, et saadavat hulka infot konsolideerida ning kuvada ühe ettevõtte kohta käiv info nii, et inimesel, kes antud süsteemi kasutab, kuluks võimalikult vähe aega võimalikult suure infohulga saamiseks. Lisaks on ettevõtte nimesid teades võimalik infot grupeerida mitte ainult ühe ettevõtte kaupa, vaid ka näiteks majandusharude kaupa.

Teiseks on meil vaja teada ka kindlasti teksti kohta käivad märksõnu. Ühe pikema uudise kirjeldamine umbes kümne märksõnaga annaks inimesele piisavalt hea arusaamise, mille kohta uudis käib ja võimaldaks saadud info põhjal vajalikke otsuseid teha. Kui polaarsus on arvutatud ja see ei lähe uudise tegeliku sisuga kokku, siis on märksõnad inimesele üheks heaks viiteks selle kohta. Ehk kui süsteemis ei ole võimalik välistada vigu, siis on lisainfo abil võimalik inimesel sellest lihtsamini aru saada ning kahtluse korral uudis ise läbi lugeda. Samas, kui analüüsitavate allikate kogus on piisavalt suur ja eeldada, et uudised on kirjutatud kasutades erinevat sõnastust, võiks minna grupeeritud andmete keskmine tulemus ikkagi kokku inimese arvamusega.

Kolmandaks tasub vaadata ka enimkasutatud sõnu tekstis. Sõnade nimekiri ja nende arv tekstis ei tarvitse küll inimesele piisavat lisainfot anda, kuid nende andmete abil võib olla võimalik teha lisaanalüüsi, mis võib, aga ei pruugi aidata muuta süsteemi uudise tonaalsuse arvutamise täpsemaks.

4.1 Kuidas leida ettevõtte nimi tekstist

Nagu eelnevalt kirjutati, on ettevõtte nime leidmine tekstist üks tähtsamaid infoühikuid tulevase teenuse jaoks. See võimaldab ühe ettevõtte või ettevõtte tegevussektori kaupa infot konsolideerida. Kuivõrd teenuse eesmärk on jälgida kümneid maailma uudisteportaale, siis ei ole eesmärk 100-protsendiline täpsus. Piisab, kui kümne uudise analüüsimisel leitaks üheksal puhul ettevõtte nimi.

Enamasti kirjutatakse uudise pealkirja ka organisatsiooni või isiku nimi, mille kohta uudis käib. Seega eeldame, et ettevõtte nime otsimiseks piisab ainult pealkirja vaatamisest. Uudiseid, mis käivad mõne ettevõtte kohta, kuid mille nimi ei esine pealkirjas, me ei vaata. Selle põhjuseks on, et sellised uudised on selgelt vähemuses. Lisaks oleks teksti analüüs, leidmaks, mille kohta uudises kirjutatakse omaette ääretult keeruline ülessanne. Seda põhjusel, et tekstist võib leida paljude erinevate ettevõtete nimesid, mis oleks näiteks konkurentide nimed või võib uudis võrrelda erinevaid ettevõtteid. Selliseid stsenaariume on palju, kus tekstis esineb rohkem kui ühe ettevõtte nimi. Uudiste pealkirjas on selliseid juhtumeid aga palju vähem ning isegi, kui selline juhtum on, saab sellest infost oma järeldusi teha, mida selles töös esialgu ei vaadelda.

Kuna antud töös on juba kasutusel *Stanford NLP*, siis tasub vaadata, mis sama raamistiku juures veel lisana kasutusel on, mida kasutada saab. Raamistikus olemasolev *Stanford Named Entity Recognizer* ehk Stanfordi nime üksuse tuvastaja märgendab sõnad vastavalt nende tüübile. Meid huvitab sõna tüüp *organization* ehk organisatsioon, mis on üks võimalikke märgendeid.

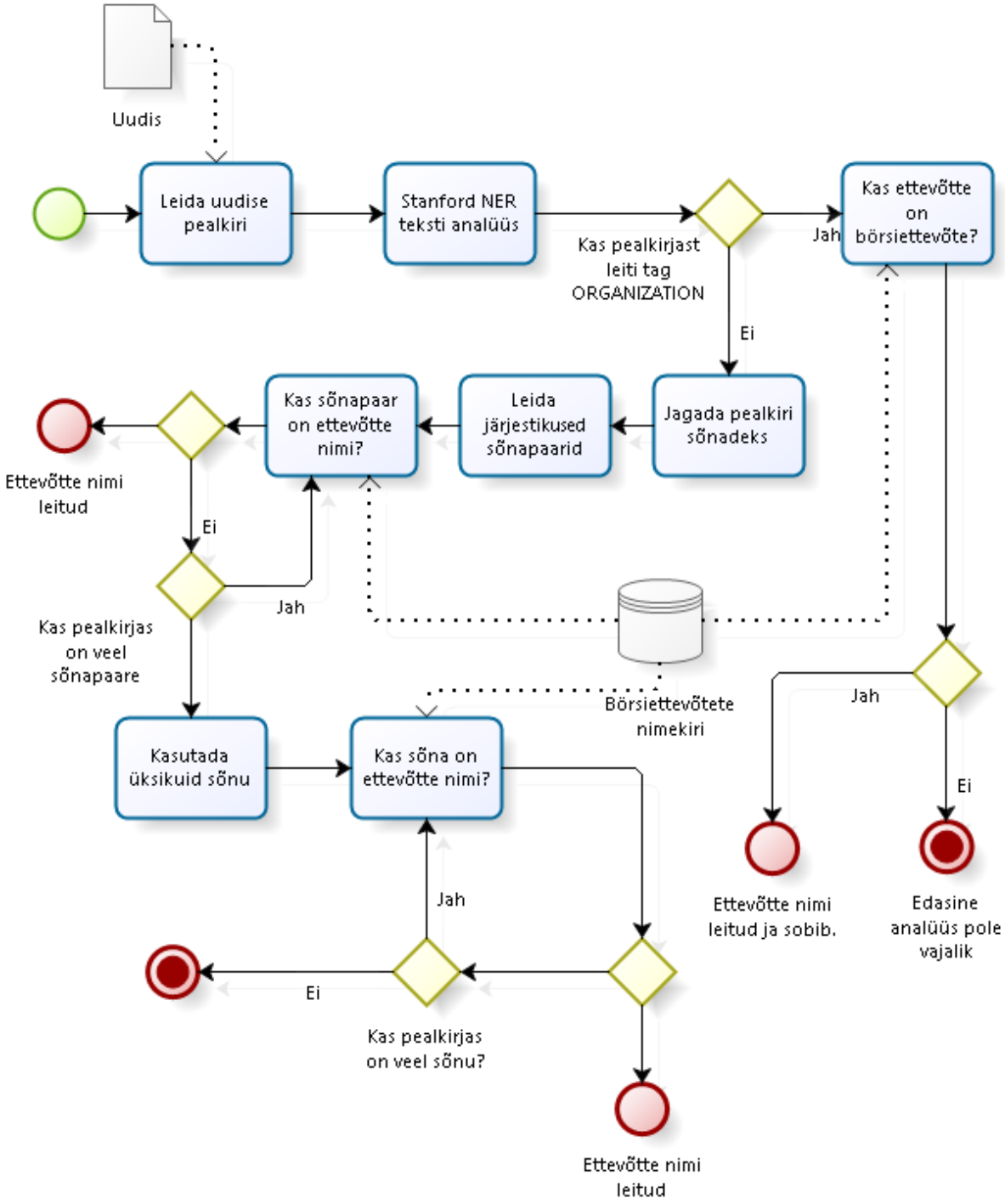
Esimese katsetusena uurime, kas *Stanford NER* suudab tuvastada pealkirjadest organisatsioonid. Katsetuse jaoks võtame veebilehelt Business Insider [28] esimesed viisteist uudise pealkirja, mis sisaldavad ettevõtte nime. Vaatame, kui suur osa uudise pealkirjades olevatest ettevõtetest tuvastati ning kas piisab ainult sellel viisil ettevõtte nime otsimisest.

Tabel 2 - Stanford Named Entity Tagger tulemused

Pealkiri	Ettevõte Pealkirjas	Tuvastati kui organisatsioon?
Avon is going wild after a report that it may sell its North America business	Avon	Ei
Ford is about to reveal a huge investment in Mexico	Ford	Jah
Athenahealth CEO said he pulls numbers 'out of his a--' in a video, and David Einhorn put it in a devastating presentation	Athenahealth	Ei
BlackRock CEO Larry Fink just told the world's biggest business leaders to stop worrying about short-term results	BlackRock	Jah
LVMH is cannibalising itself in China	LVMH	Jah
Deutsche Bank made something funny to compete with The Economist's 'Big Mac Index'	Deutsche Bank	Jah
Zillow just slashed its revenue outlook	Zillow	Ei
Goldman Sachs is a tech company	Goldman Sachs	Jah
GOLDMAN SACHS PRESIDENT: A teacher told my parents if they were really lucky I might grow up to be a truck driver	Goldman Sachs	Ei
Wells Fargo beats estimates	Wells Fargo	Ei
Johnson & Johnson is still getting whacked by the strong dollar	Johnson & Johnson	Jah
CREDIT SUISSE CEO: The pecking order has changed on Wall Street	Credit Suisse	Ei
The Wells Fargo employee who emailed the CEO asking for a $10K raise for his colleagues has quit	Wells Fargo	Ei
Tesla's new gigafactory will highlight the 2 biggest labor trends in America	Tesla	Ei
GE's historic deal was a huge win for shareholders ... at the expense of GE's creditors	GE	Jah

Selle tulemusena näeme, et *Stanford NER* tuvastas viieteistkümnest organisatsioonist seitse (vt Tabel 2 - Stanford Named Entity Tagger tulemused). Seega võime kohe öelda, et sellest üksi ei piisa ettevõtte nime leidmiseks. Samas on selle kasutamine esmase kontrollina siiski mõistlik.

Järgmise sammuna kontrollime pealkirjas olevaid sõnu börsiettevõtete nimekirjaga võrreldes. Esimesena tasub kontrollida järjestikku asuvaid sõnu ning vaadata, kas mõni sõnapaar vastab ettevõtte nimele, mis koosneb kahest sõnast. Teise sammuna tuleb kontrollida sõnu eraldi, ning leida, kas see vastab mõnele ettevõtte nimele. Sellise järjekorraga välistame võimalused, kus ettevõtte nimes sisaldub mõne teise ettevõtte nimi ning seetõttu leitakse pealkirjast vale ettevõtte nimi.



Joonis 3 - Ettevõtte nime leidmine

4.2 Märksõnad tekstist

Märksõnade leidmiseks kasutame Maui-indexer'i versiooni 1.2. Tegemist on projektiga, mis on mõeldud märksõnade leidmiseks tekstist. Olena Medelyan'i doktoritöö, mille üheks tulemuseks oli antud tarkvara, väitel [29] on maui-indexer raamistik võimeline teksti teemade märksõnu välja tooma inimesega võrreldaval tasemel või olla antud ülessandes inimesest parem. Seega märksõnade leidmiseks ei hakata selle töö raames ühtegi uut meetodit looma ja kasutatakse olemasolevat tarkvara.

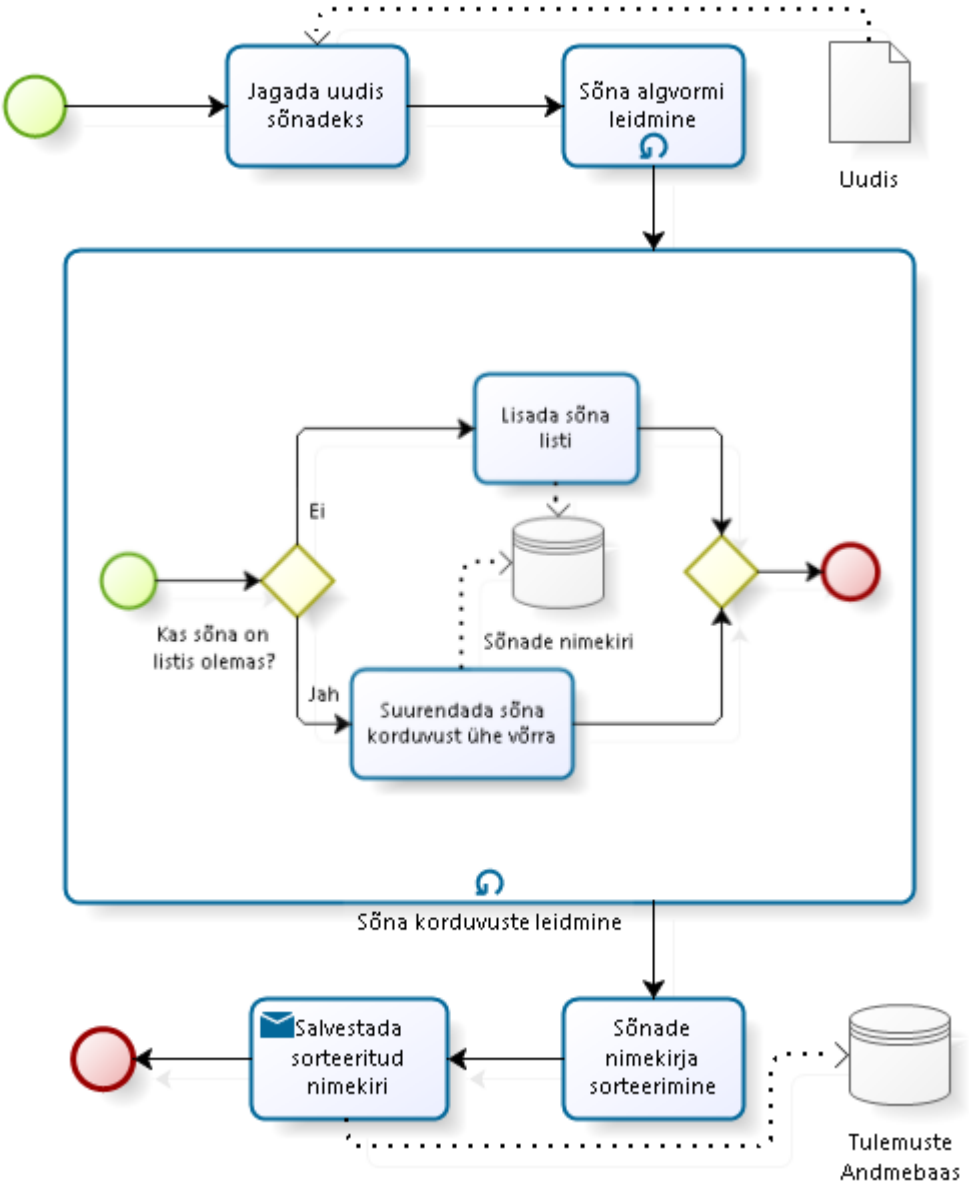
Töös kasutame kasutame maui indexer'i raamistiku näidiskoodi muutmata kujul ilma kindla sõnastikuta.

4.3 Enimkasutatud sõnad tekstis

Enimkasutatud sõnade leidmiseks pole vajalik väliseid raamistikke kasutama hakata, kuna lahendust luues on võimalik veenduda, et seal ei ole midagi üleliigset, mis võib lahendust kohmakamaks muuta.

Välistamaks sama sõna erinevatel kujudel esinemist teisendame tekstis leitavad sõnad algvormi kasutades Stanford NLP raamistikku, mida juba teised töö osad kasutavad.

Pärast sõna teisendusi lisatakse sõna Java *Map*'i, kus võtmeks on sõna ise ning väärtuseks sõna korduvuste arv tekstis, mille väärtust suurendatakse ühe võrra iga kord, kui sama vaste Java *Map*'ist leitakse, vastasel juhul lisatakse sinna uus väärtuste paar, mille võtmeks on sõna ja väärtuseks üks.



Joonis 4 - Enimkasutatud sõnade leidmise protsess

5. Tulemuste hindamine

Töös on eelnevalt leitud tulemused teksti sisu kohta kasutades kahte erinevat leksikaalset ressurssi, mida on ka minevikus kasutatud teiste tööde juures. Selle töö üheks eesmärgiks on leida uus viis tulemuste kombineerimiseks ning kontrollida, kas see on täpsem kui eelnevalt kasutatud meetodid eraldi. Peale arvutatud tulemuste on vaja analüüsiks teada ka inimese hinnangut artikli sisu kohta.

Tulemusi hindab esimesena inimene, kes annab hinnangu, kas uudise sisu on tema arvates positiivse tooniga, neutraalse tooniga või negatiivse tooniga antud ettevõtte jaoks, mille kohta uudis käib. Seejärel hindab samade uudiste sisu kolmandas peatükis kirjeldatud süsteem ning seejärel võrdleme, kas arvuti poolt antud tulemuste vahel on olemas mingisugune korrelatsioon.

Inimese poolt antud uudise polaarsus on defineeritud kolme diskreetse väärtusena, milleks on positiivne, neutraalne ja negatiivne.

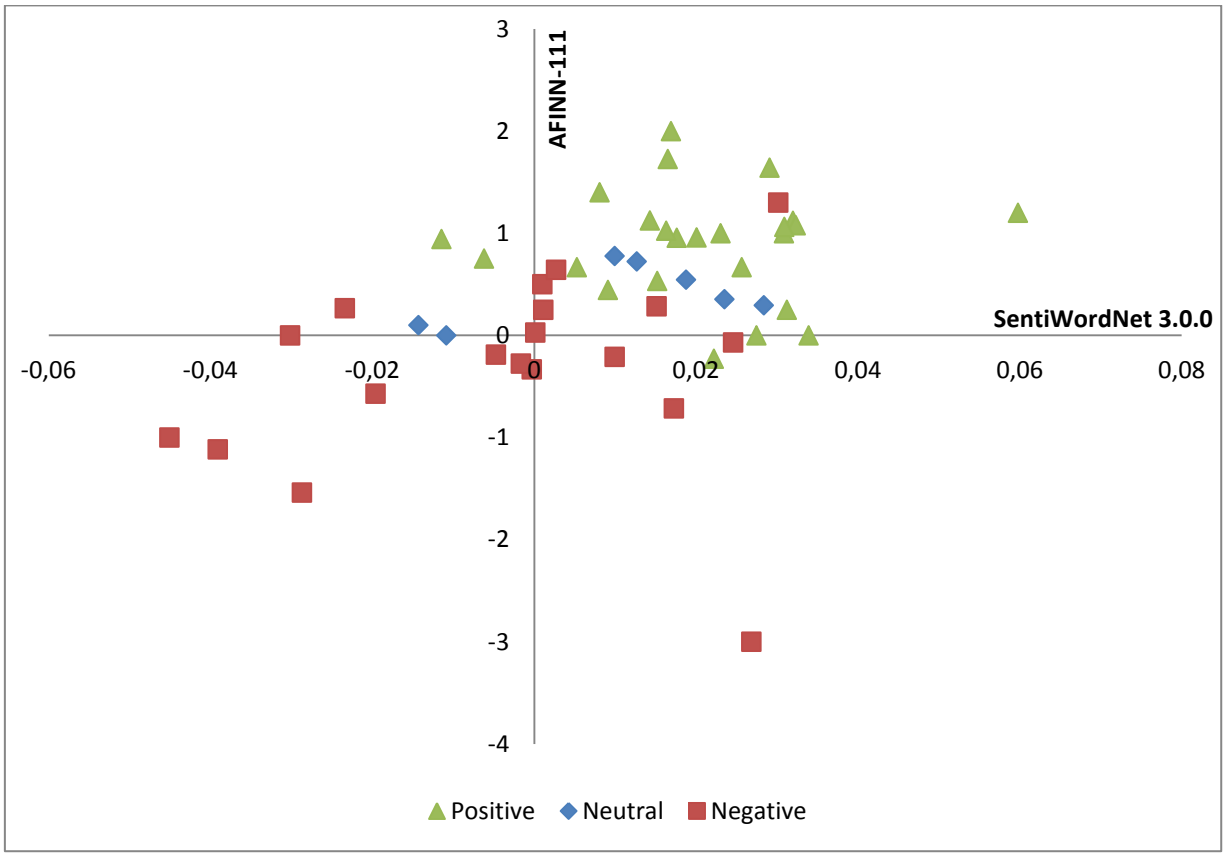
Kokku valiti töö jaoks viiskümmend uudist viiest erinevast uudisteportaalist, kus igast uudiste allikast valiti võrdne arv uudiseid.

5.1 Meetodi valik

Kuna eelneva analüüsi tulemusi on võimalik kasutada mitmel erineval viisil, siis uurime, milline viis on meile kõige sobivam ning annab parima tulemuse. Iga uudise kohta on teada järgnevad andmed: inimese antud hinnang, SentiWordNet 3.0.0 andmebaasist leitud sõnade hulk ning sõnade tonaalsuste summa ja AFINN-111 andmebaasist leitud sõnade hulk ning nende sõnade tonaalsuste summa.

Meetodid, mida on võimalik nende andmete abil koostada on saadud sõnade keskmine ja sõnade summa ilma sõnade arvu arvestamata. Lisaks on võimalik kombineerida mõlemat meetodit ja saada kolmas meetod tulemuste arvutamiseks. Esiteks vaatleme sõnade keskmist, teiseks vaatleme sõnade summasid. Viimasena vaatleme kahe eelneva meetodi keskmist.

Esimesel juhul on kummalgi juhul sõna tonaalsus T arvutatud valemiga $T = \frac{\sum S_{ton}}{S_{Arv}}$, ehk tegemist on sõnade keskmisega. Edaspidi tähistab S_{ton} sõnade tonaalsuse summat ning S_{Arv} on leitud sõnade arv tekstis.



Joonis 5 - Analüüsi tulemused (sõnade keskmine)

Sellisel juhul on gaafikult visuaalselt näha, et tõesti on olemas mingisugune korrelatsioon teksti polaarsuse ja teksti asukoha vahel graafikul. Kohe hakkab silma, et kui mõlema telje väärtus on negatiivne, on ka uudis negatiivse polaarsusega. Samas kui mõlemad väärtused on positiivsed, ei pruugi uudis olla alati positiivse tooniga.

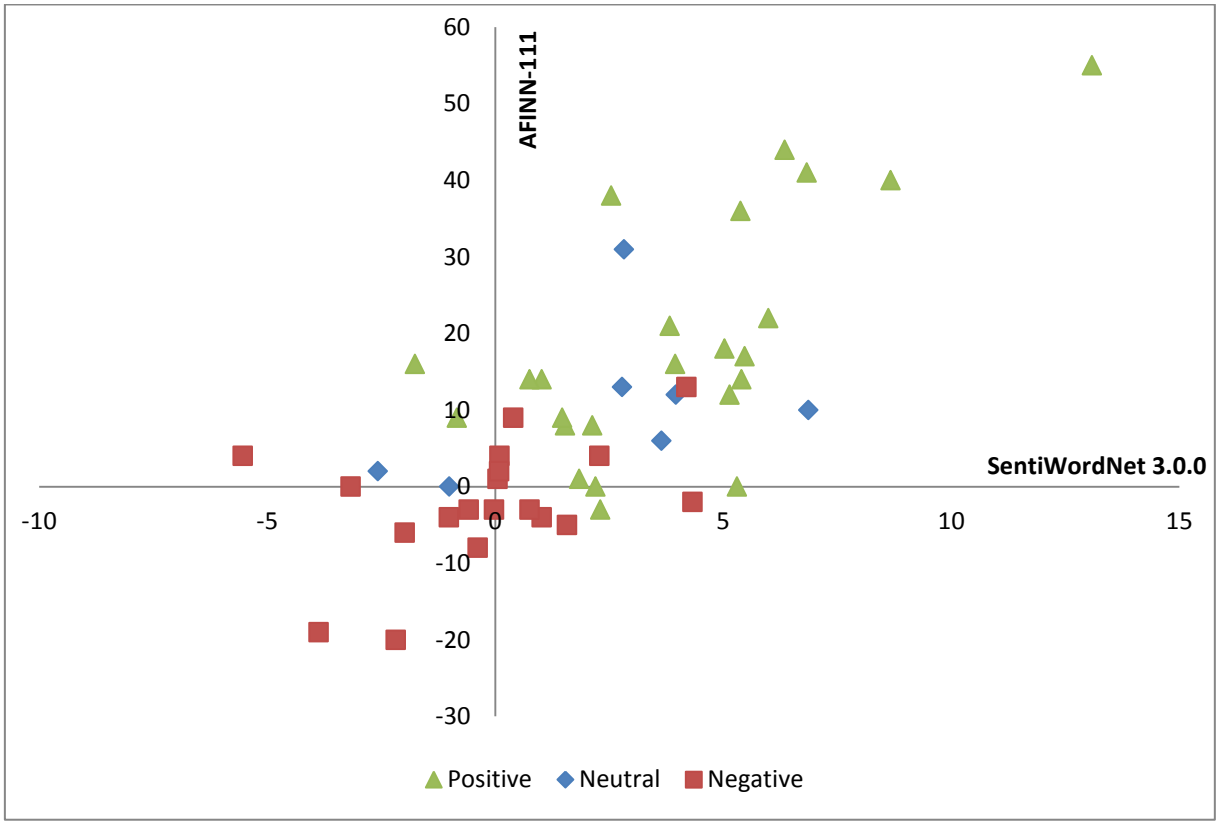
Lisaks on näha, et nullist positiivses suunas üks kaugemaid punkte on negatiivse sisuga uudis ehk tegemist on veaga teksti analüüsis. Antud uudise puhul on tegemist uudisega, mille sisuks on, et PB kasum kukkus 19% nafta hinna languse tõttu. [30] Seega pole võimalik välistada, et mõni uudis on antud analüüsi meetodi järgi hoopis teise tonaalsusega kui inimese arvates uudis seda on.

Ka on näha, et kui tõmmata sirge joon läbi graafiku viisil, et joonest üles jääks võimalikult palju positiivseid uudiseid ning alla võimalikult palju negatiivseid uudiseid, siis sellisel juhul

jääks positiivsete sekka kolm negatiivse tooniga uudist ning negatiivsete piirkonda kaks positiivse tooniga uudist.

Lisaks on näha, et neutraalse sisuga uudised oleks eelnevalt tõmmatud joonest sarnasel kaugusel ning enamus positiivseid uudiseid oleks kujutletavast joonest kaugemal.

Teisel juhul võtame vaatluse alla sõnade summa, kus arvutame teksti tonaalsust kummalgi juhul T valemiga: $T = \sum S_{ton}$



Joonis 6 - Analüüsi tulemused (sõnade summa)

Teisel juhul näeme, et uudis, mis ei lähe oma sisu ja asukoha poolest omavahel kokku, on tulnud lähemale nullpunktile ning sellest on möödunud suur osa positiivseid uudiseid. Seega ilma sõnade arvu arvestamata oleks tegemist antud uudise puhul siiski positiivse uudisega arvuti analüüsi järgi, kuid viga oleks väiksem kui eelneval juhul.

Kui tõmmata joon nii, et joonest alla jääks võimalikult palju negatiivseid uudiseid ning ülespoole võimalikult palju positiivseid uudiseid, siis oleks tulemus sama, mis eelmisel asukoha arvutamise viisil. Joonest alla jääks kaks positiivset uudist ning üles kolm negatiivset uudist.

Samas on sellel viisil näha, et neutraalse sisuga uudised on rohkem laiali hajunud, kui eelmisel juhul. Lisaks on näha, et mitmed positiivse sisuga uudised on tulnud lähemale kujutletavale joonele, mis peaks eraldama positiivseid uudiseid negatiivsetest uudistest.

Seega on mõlemal viisil oma häid ja halbu külgi. Kindlasti tasub uurida, mis mõlema meetodi keskmine tulemuseks annab.

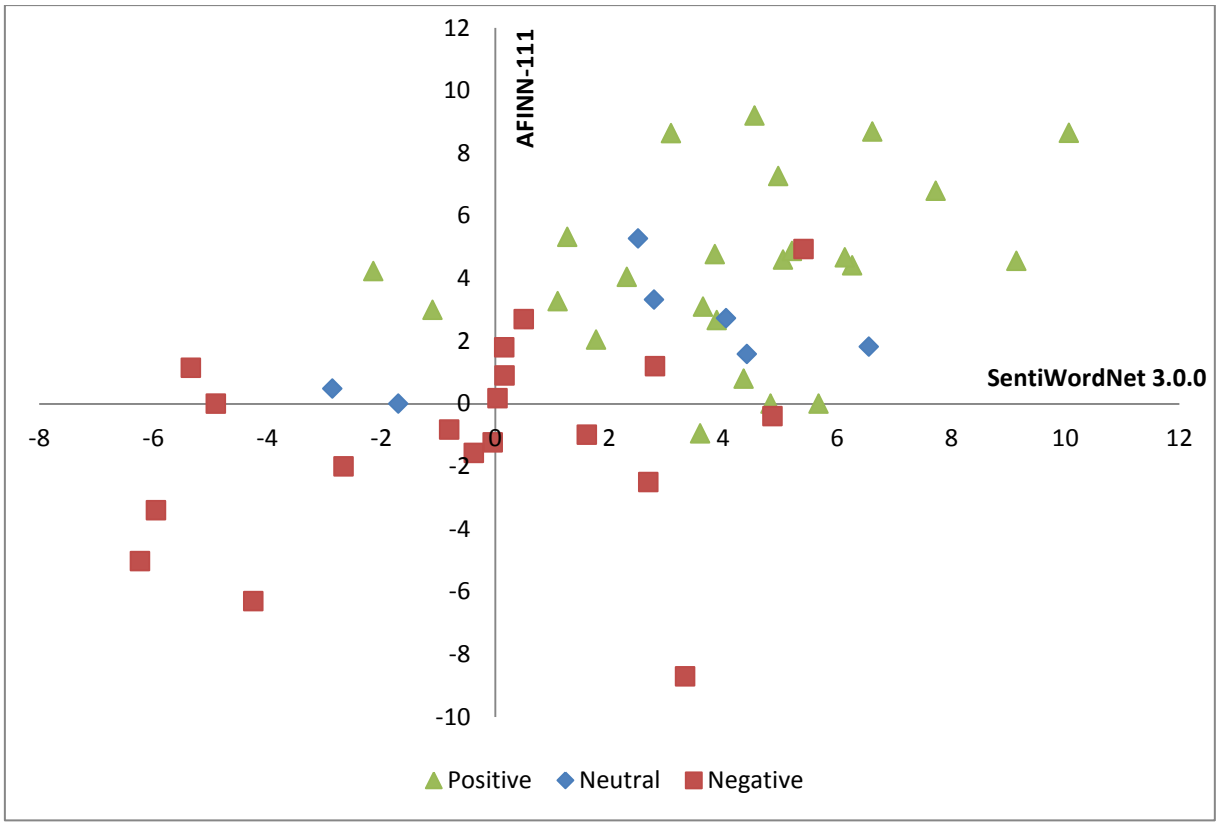
Kuna kolmanda meetodi jaoks oleks vaja, et meetodite osatähtsus uues meetodis oleks võrdne, siis lisame keskmise arvutamisel väiksemale väärtusele ehk sõna keskmise meetodi väärtusele kordaja.

SentiWordNet 3.0.0 kordaja oleks kõige kaugema punkti järgi võttes 13,09/0,0598 = 218,89, mille ümmardame 220'le

AFINN-111 kordaja oleks 55/2 = 27,5. Antud väärtuse ümmardame 28 peale. Kuna AFINN-111 kahe meetodi keskmise summa oleks umbes viis korda suurem kui SentiWordNet 3.0.0 meetodite keskmine, siis korrutame jagaja viiega. Selle eesmärk on, et telgede pikkus oleks sarnane ning sellisel juhul oleks ka sõna üldise tonaalsuse arvutamine kahe telje puhul lihtsam. Seega tonaalsuse arvutamise valemid on järgnevad:

$$T_{AFINN} = \left(\sum S_{ton} + \frac{\sum S_{ton}}{S_{Arv}} * 28\right)/2$$

$$T_{Swn3} = \left(\sum S_{ton} + \frac{\sum S_{ton}}{S_{Arv}} * 220\right)/10$$



Joonis 7 - Analüüsi tulemused (kahe meetodi keskmine)

Kolmandal juhul on kohe näha, et negatiivne uudis, mis sisult kokku ei sobi antud meetodiga on nullpunktist kaugusel, kus sellest kaugemale jääb sarnane hulk uudiseid nagu teises meetodis ehk see sobib meile paremini kui esimene meetod.

Sellel juhul on veel näha, et neutraalsed uudised on vähem laiali hajutatud kui teisel juhul ning nende kaugus nullpunktist on enam-vähem sarnane ehk neutraalsete asukoht on lähemal esimesele meetodile.

Lisaks on selle meetodi puhul näha, et kui tõmmata joon nii, et joonest alla jääks võimalikult palju negatiivseid uudiseid ning ülespoole võimalikult palju positiivseid uudiseid, siis oleks positiivsed uudised sellest kujutletavast joonest kaugemal.

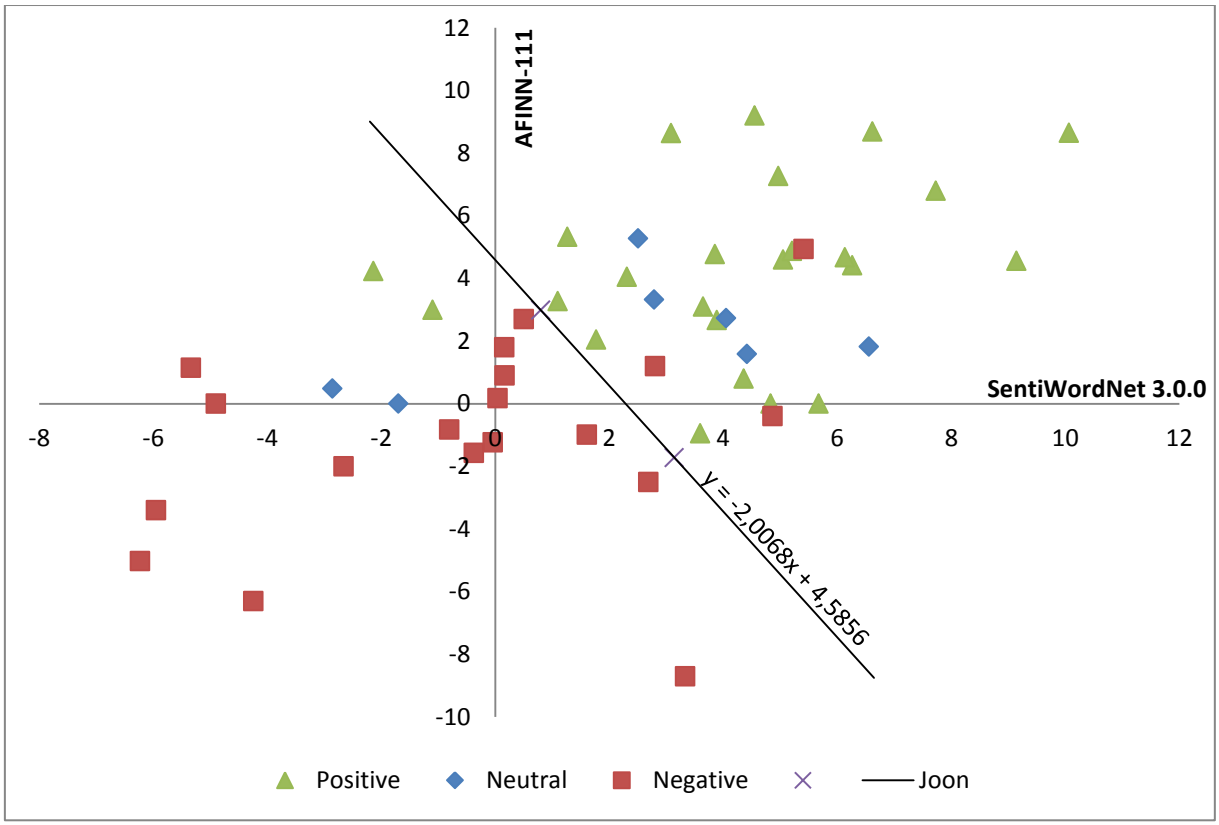
Ka sellise jaotuse puhul oleks negatiivsete uudiste arv joonest üleval kolm ja positiivsete uudiste arv joonest all kaks.

Kuna kolmas meetod tundub esmasel vaatlusel toimivat kõige paremini, siis kasutame edasiseks analüüsiks seda meetodit.

5.2 Tonaalsuse leidmine

Esimeseks kõige lihtsamaks analüüsi viisiks proovime leida kaks punkti, mida joon läbida võiks ning arvutada uudise kauguse sellest joonest. Joone leidmiseks leiame kaks negatiivse ja positiivse uudise paari, mis asuvad lähestikku ning asuvad nii, et loodav joon jaotaks kaardi nii kaheks, et ühele poole jääks võimalikult palju positiivseid ja teisele poole võimalikult palju negatiivseid uudiseid.

Sellise analüüsi tulemuseks oleks tonaalsus, mis on defineeritud viisil, et kõiki väärtuseid, mis on nulliga võrdsed või suuremad, loetakse positiivseteks uudisteks ning kõiki negatiivseid väärtuseid loetakse negatiivse tooniga uudisteks.



Joonis 8 - Esimene tonaalsuste jagamise meetod

Sellise meetodi puhul saame ühe joone, mis jagab graafiku kaheks osaks. Esiteks on see hea põhjusel, et sellise meetodiga saab anda ühele punktile ühe numbrilise väärtuse, milleks oleks punkti (x;y) kaugus joonest (Ax + By + C = 0) ning siis vastavalt kas positiivne või negatiivne kordaja saadud väärtusele olenevalt, kas punkt asub joonest vasakul või paremal.

$$d = \frac{|Am + Bn + C|}{\sqrt{A^2 + B^2}}$$

Sellisel puhul on näha, et valele poole joont jääb kaks positiivset ja kolm negatiivset väärtust. (vt. Lisa 5) Seega on selgelt valed viis väärtust viiekümnest. See tähendab kindlalt, et valed on 10% väärtustest. Selline veaprotsent pole küll ideaalne, kuid on loodava süsteemi jaoks veel aktsepteeritav. Lisaks võib eeldada, et juhul kui tuleb välja mõni tähtis uudis, siis on erinevate autorite poolt kirjutatud uudis sama sündmuse kohta korraga rohkem kui ühes uudisteportaalis. See tähendaks, et kasutajale kuvatakse ikkagi pärast andmete koondamist õige tonaalsusega info vaadeldaval päeval mingi ettevõtte kohta.

5.3 Tulemuste võrdlus

Saime teada, et arvutades kauguse loodud piirist, on sellisel juhul uudise tonaalsuse analüüsi puhul valesti arvutatud 10% uudistest. Kuid seda tasub võrrelda nii SentiWordNet 3.0.0 kui ka AFINN-111 saadud tulemuste täpsusega.

Vaatleme kumbagi meetodit eraldi kõigil kolmel erineval juhul. Esiteks kõikide sõnade summana, teiseks sõnade summa jagatud sõnade arvuga ning viimaks ka arvutatud kahe meetodi keskmist. Kõikidel juhtudel leiame piiri, kus on kõige vähem vigu ehk juhu, kus kummalegi poole piiri jäävate vale tonaalsusega uudiste arv on väikseim. Valeks tonaalsuseks loeme juhte, kus on positiivsel alal negatiivse tonaalsusega uudis või negatiivse tonaalsusega alal positiivne uudis.

Tabel 3 - Meetodite vea protsent

	Sõnade summa	**Sõnade keskmine**	**Kahe meetodi keskmine**
SentiWordNet 3.0.0	16%	16%	16%
AFINN-111	12%	14%	12%

Sellest tulenevalt võime järeldada, et kahe erineva meetodi abil leitud uudise tonaalsuse leidmine on täpsem kui meetodid eraldi võttes on.

Lisaks näeme, et uudiste analüüsi puhul on AFINN-111 täpsem kui SentiWordNet 3.0.0.

6. Mittekvantitatiivsete näitajate seos tonaalsusega

Kuna eelnevalt loodud tonaalsuse leidmise meetodi puhul tekib viga kümnel protsendil analüüsitud juhtudest, siis tasub edasi uurida, kas eelnevalt välja toodud mittekvantitatiivsete näitajate ja teksti vahel on mõni seos, mida saaks kasutada eelneva meetodi parendamiseks.

6.1 Märksõnad ja tonaalsus

Et leida, kas on olemas seosed uudiste märksõnade ja polaarsuste vahel, siis esialgu leiame prooviks eelnevalt kasutatud uudiste hulgast kümme selgelt positiivselt ja kümme selgelt negatiivset uudist ning hindame saadud märksõnu ühe suure grupina, kasutades kolmandas peatükis kirjeldatud polaarsuse leidmise süsteemi.

Analüüsides positiivseid sõnu tuleb SentiWordNet 3.0.0 sõnade summaks 0.9815 ning tabelist leiti 52 sõna. AFINN-111 puhul leiti 9 sõna, kus sõnade polaarsuste summa oli 8.

Negatiivsete uudiste märksõnu uurides leiti SentiWordNet 3.0.0 tabelist 43 sõna, mille polaarsuste summa oli -2.3765 ning AFINN-111 puhul leiti 7 sõna polaarsuste summaga -8.

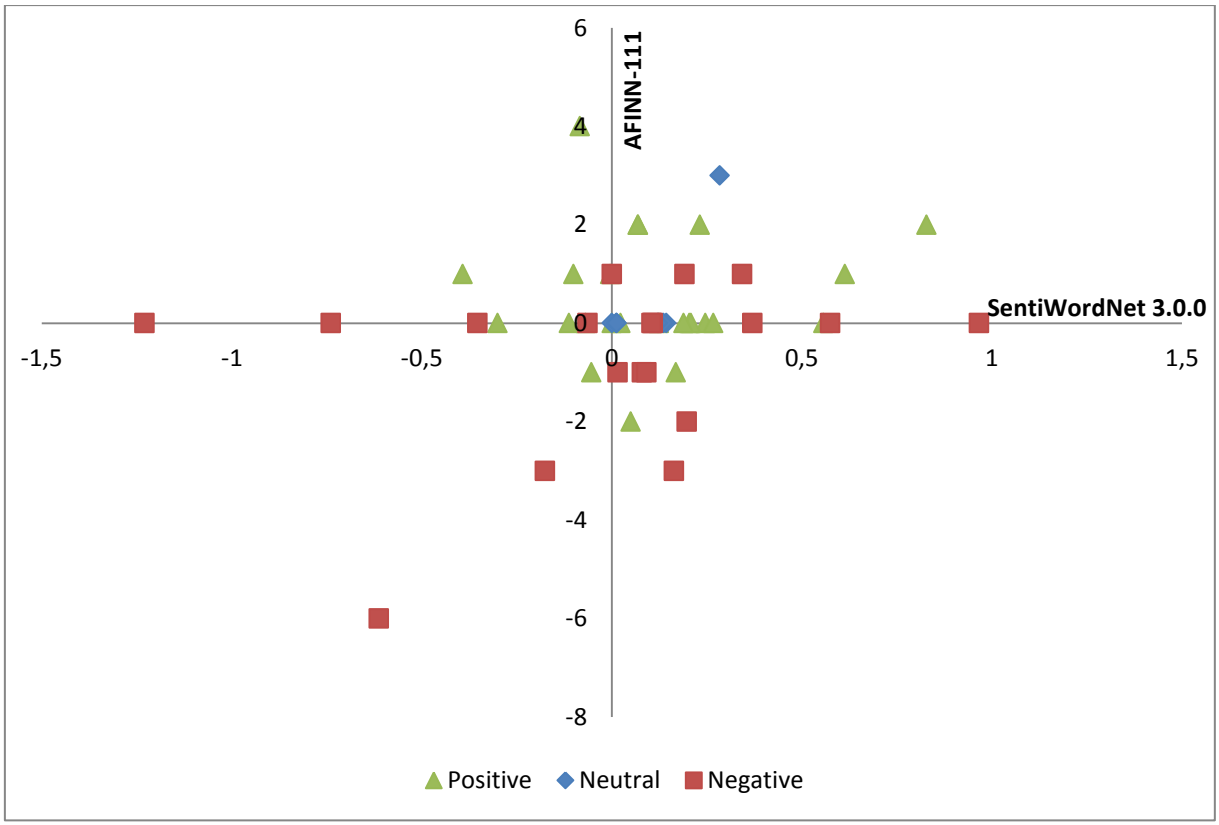
Seega võib eeldada, et ka märksõnade polaarsuste ja uudise sisu vahel on olemas seos. Lisaks otsime, kas uudise puhul, kus on tekkind vastuolu uudise sisu ja leitud teksti polaarsuste vahel, on võimalik teha märksõnade abil mingeid muid järeldusi.

Tabel 4 - Vastuoluliste uudiste märksõnade tonaalsus

Pealkiri	Polaarsus	SWN3 Skoor	AFINN Skoor
BP profit hit by oil price fall	Neg	0,3705	0
Microsoft: The post-Windows company?	Pos	-0,3004	0
Deutsche Bank: This Week's Tesla Announcement Could Be a Bigger Deal Than Investors Realize	Pos	0,2461	0
Xerox cuts 2015 profit forecast due to strong dollar	Neg	0,1976	-2
Credit Suisse CEO says post-crisis expansion was a mistake	Neg	0,1635	-3

Vastuoluliste uudiste puhul on näha, et kolmel juhul viiest võiks see mõjutada uudise polaarsust suunas, mis vastaks ka inimese arvamusele.

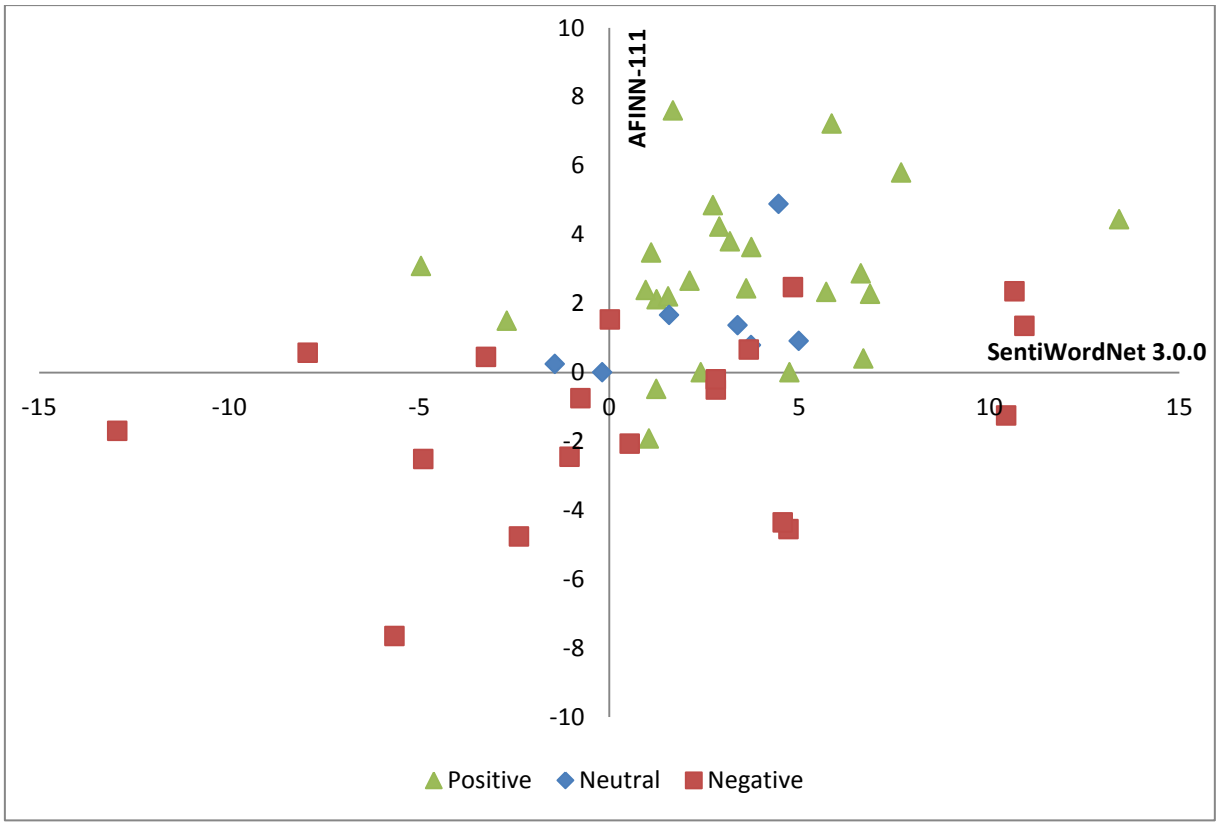
Kõigepealt uurime kõikide uudiste märksõnade tonaalsusi.



Joonis 9 - Märksõnade tonaalsus (sõnade summa)

Nagu siit jooniselt näha on, ei ole siis esialgne pilt paljulubav, samas on seda joonist vaadeldes võimalik välja tuua mingid piirkonnad, kus on rohkem positiivseid uudiseid või rohkem negatiivseid uudiseid. Lisaks on näha, et üheks probleemiks antud juhul võib olla ka osade väärtuste nulli jäämine. Seda eriti AFINN-111 andmebaasi kasutades, mis on arvatavasti nii selle tõttu, et seal andmebaasis on umbes 40 korda vähem ridu.

Seega proovime edasi kasutades polaarsuse kahe meetodi keskmist ning võtame seejärel märksõnade saadud tulemuse ning teksti polaarsuse sellel viisil saadud polaarsuste keskmise ning kujutame ka seda graafikul.



Joonis 10 - Teksti ja märksõnade arvutatud keskmised

Juba seda graafikut vaadeldes võime väita, et kuigi alguses tundus, et märksõnade ja teksti polaarsuse vahel on seos olemas, siis üksikuid uudiseid arvesse võttes muudab märksõnade kasutamine eelnevat teksti analüüsi ebatäpsemaks.

Prooviti veel viisi, kus vähendati märksõnade kaalu, kuid ka sellisel puhul oli kohe selgelt näha, et kõikidel juhtudel on see ebatäpsem kui esialgne ilma märksõnadeta kaart. Seega ei ole mõistlik märksõnade ja teksti polaarsuse vahelisi seoseid edasi vaatlema hakata.

6.2 Enimkasutatud sõnad ja tonaalsus

Lisaks tasub uurida, kas on olemas seoseid enimkasutatud sõnade ja teksti tonaalsuse vahel. Selleks uurime kõigepealt uudiseid, mis on nii inimese arvates kui ka eelnevalt loodud tonaalsuse analüüsi alusel negatiivse sisuga. Edasi vaatame, kas neid seoseid tasub edasi uurida või mitte.

Antud juhul vaatleme sõnu, mida oli tekstis rohkem kui korra ning analüüsime nende sõnade polaarsusi. Sõnade arvu antud juhul ei arvestata ehk iga sõna polaarsust kasutatakse ainult korra ning seda ei korrutata läbi selle esinemise arvuga tekstis, kuna sellisel juhul oleks tegemist sama analüüsimeetodiga, mis terve teksti analüüs eelnevalt oli.

Tabel 5 - Enimkasutatud sõnade tonaalsuste summa

Pealkiri	SentiWordNet Skoor
KFC 'dreadful,' says man who brought franchise to Britain	-0,105353768
Google Looks to Buy Patents in Experiment to Lessen Lawsuits	0,013734837
Deutsche Bank shares fall on shake-up plan	-1,183640574
Ellen Pao faces $1m legal bill in sexism case	-0,314924432
Deutsche Bank first-quarter profit falls by half as legal charges bite	0,152260042

Juba viie esimese uudise analüüsil hakkab silma, et sellel meetodil on samu vigu, mis märksõnade ja polaarsuste seoste leidmisel. Esiteks on kohe esimesel viiel juhul AFINN-111 andmebaasi kasutades leitud sõnade arv null. Teiseks probleemiks on, et juba viit uudist analüüsides on näha, et SentiWordNet 3.0.0 andmebaasi abil arvutatud sõnade skoor on kahel juhul vale. Sellest võib eeldada, et kuigi antud juhul võib mingeid seoseid antud meetodi ja teksti reaalse sisu vahel leida, ei ole see täpsem kui oli märksõnade leidmine.

Kokkuvõtteks võib väita, et kuigi märksõnadest võib inimese jaoks tekstist arusaamiseks kasu olla (Vt Lisa 3 ja 4), pole märksõnade ja enimkasutatud sõnade abil võimalik teksti polaarsuse analüüsi täpsemaks teha. Pigem vastupidi, sest nende sõnade kasutamine muudaks analüüsi tulemusi ebatäpsemaks.

7. Kokkuvõte

Töö põhieesmärgiks oli leida meetod uudise teksti tonaalsuse hindamiseks. Selleks kasutati kahte sõnakogumit, mis on mõlemad mõeldud teksti tonaalsuse leidmiseks. Lisaks oli väga tähtis leida ettevõtte nimi teksti pealkirjast, mis lahendati kasutades *Stanford NER* raamistikku ja börsiettevõtete nimekirja. Viimaseks oli vaja leida tähtsamad märksõnad igast uudise tekstist. Selleks kasutati muutmata kujul maui-indexer'i raamistikku, mis võimaldab antud ülesannet lahendada inimesega võrdsel tasemel.

Töö oluliseima tulemusena on loodud uudiste analüüsi meetod, mille tulemusena saab teada uudise tonaalsuse suurema täpsusega kui kumbagi olemasolevat sõnakogumit SentiWordNet ja AFINN eraldi kasutades.

Kahe erineva leksikaalsete teadmiste kogumiga on võimalik leida teksti tonaalsus täpsemini kui ühte kogumit kasutades. Vaadeldes kumbagi sõnakogumit eraldi kasutades saadud tulemusi, oleksid need tulemused vähem täpsed kui loodud meetodi korral. Samas pole vigade välistamine võimalik ja ka sellisel juhul on kümme protsenti uudistest tegelikkuses vastupidise tonaalsusega kui on tulemus arvutuslikult. SentiWordNet 3.0.0 puhul oleks parimal juhul 16% uudiseid vastupidise tonaalsusega. AFINN-111 puhul oleks parimal juhul 12% uudiseid vastupidise tonaalsusega.

Lisaks võib väita, et loodud meetodit kasutades pole võimalik määrata ühte neutraalset piirkonda. Seega seda antud töös ei tehtud. Saadud tulemused on kas positiivses või negatiivses alas ning number kirjeldab kui positiivne või negatiivne antud uudis analüüsi tulemuse järgi on.

Edasiseks uurimiseks tasuks vaadelda teksti tonaalsuse leidmist ka *naive* Bayesi meetodit kasutades, mida võrreldi Kyle Thompson ja Nilayan Bhattacharya töös [7], nii SentWordNeti kui AFINN'i kasutamisega. Antud meetodi abil on võimalik teksti tonaalsuse leidmisel saavutada vähemalt 80% täpsus [31] ja isegi kuni 90% täpsus. [9]

Lisaks tasub edaspidi kaaluda ühe ühtse tulemuse leidmisel peale arvutusliku meetodi ka mõnda masinõppe süsteemi. Sellisel juhul võib, aga ei pruugi tulla täpsem tulemus, kuid samas on raskem määratleda, miks see konkreetne tulemus saadi.

1. Kas eesmärk saavutati?

Kõige tähtsamaks eesmärgiks antud töös oli leida meetod, mis suudab analüüsida suurt hulka uudiseid ning leida uudiste tonaalsus. See eesmärk täideti ning leiti meetod, mis suudab analüüsida uudiste sisu piisava kiirusega, et oleks võimalik analüüsida arvestatavat hulka uudisteportaale korraga. Lisaks on loodud süsteemi täpsus uudise tonaalsuse leidmisel piisav, et seda oleks mõistlik edasi arendada.

2. Põhitulemuste loetelu

Esimese tulemusena leiti viis, kuidas saada tekstist kahte erinevat sõnakogumit kasutades välja arvutada kummagi meetodi abil eraldiseisvad tulemused sama teksti kohta.

Nagu eelnevalt sai välja toodud, on see töö tähtsaimaks tulemuseks. Nende andmete abil loodi analüüsimeetod, mille abil saab leida teksti tonaalsuse. Lisaks toimib uus meetod paremini, kui kumbagi kasutatud sõnakogumit eraldi kasutades.

Järgmisena leiti viis tekstist märksõnade leidmiseks. Selle jaoks kasutati juba olemasolevat tekstianalüüsi raamistikku.

Viimasena oli vaja leida, missuguse ettevõtte kohta avaldatud uudis käib. Selle ülesande lahendamise jaoks kasutati Stanford NLP raamistikku ning lisaks veel olemasolevate börsiettevõtete nimekirja.

3. Kas eesmärgid saavutati?

Töö põhieesmärk oli leida meetod tekstianalüüsiks, mis ka saavutati. Sellise analüüsi meetodiga pole ilmselt võimalik tulemust enam märgatavalt täpsemaks saada ning nagu mainitud, tuleks edasiseks parendamiseks kasutada ilmselt mõnda masinõppe algoritmi.

Lisaks oli vaja välja selgitada, missuguse ettevõtte kohta analüüsitav uudis käib. Ka see eesmärk täideti ning leitud meetod toimib küllalt hästi, et seda tulevikus oleks loodava teenuse juures on võimalik kasutada.

Kolmandaks eesmärgiks oli leida ka märksõnad tekstist, mida inimesele kuvada, et ta saaks kiiremini otsuseid vastu võtta. Selleks kasutati eelnevalt loodud süsteemi, mis teeb seda piisavalt hästi, et seda kasutada tasuks ning seda parendada pole enam mõistlik.

Summary

The purpose of this thesis was to create a basis for a startup, that is capable of analyzing polarity of news in real time from selected news portals. Therefore the central problem of this work was to find polarity of a given text. Secondary objectives were finding keywords from the text and to find a way to get organization names from the news headlines.

For sentiment analysis the primary question was how to combine two lexical recourses in a way that the new method would use both resources and it would be better at classifying polarity of a given text than using those resources by themselves. Secondly for finding keywords there was found a tool which already functions as well as a person. For finding organizations from the news headlines there was used a library named Stanford NLP. However this tool finds about half of the organization names. To improve that there was proposed a simple way to find the necessary organization names, which uses a list of known company names.

Primary result of the thesis was a new approach for analyzing news polarity that combines SentiWordNet and AFINN lexical recourses into one method. That classifies 10% of the results into the wrong tonality category. However using SentiWordNet and AFINN recourses alone respectively 16% and 12% of the results were wrongly classified.

In addition to those results one can say that using this method it is possible to classify the news as either positive or negative and it is not practical to try to classify neutral news.

For finding the keywords from the news a library named Maui-indexer was used, which does find keywords on a level comparable to a human. Finally there was proposed a way to find organization names from the news headlines that works in most cases.

Kasutatud kirjandus

[1] Chan, W. S., Stock price reaction to news and no-news: drift and reversal after headlines - *Journal of Financial Economics*, 2003, 70, (2), 223–260

[2] Veronesi, P., Stock market overreactions to bad news in good times: a rational expectations equilibrium model - *Review of Financial Studies*, 1999, 12, (5), 975-1007

[3] Gidófalvi, G., Using News Articles to Predict Stock Price Movements : teadustöö aruanne. University of California, San Diego, 2003

[4] Antweiler, W., Murray, Z. F., Do US Stock Markets Typically Overreact to Corporate News Stories? 2006 [online] Social Science Research Network (13.05.2015)

[5] Ding, X., Liu, B., Yu, P. A holistic lexicon-based approach to opinion mining. - *In Proceedings of the Conference on Web Search and Web Data Mining (WSDM-2008)*, 2008. 231-240

[6] Baccianella, S., Esuli, A., Sebastiani, F. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining - *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010, 2200-2204

[7] Thompson, K., Bhattacharya, N., Opinion Mining and Name Entity Detection from News Comments : ainetöö. University of Georgia, Athens, 2014

[8] Bravo-Marquez, F., Mendoza, M., Poblete, B., Combining Strengths, Emotions and Polarities for Boosting Twitter Sentiment Analysis - *Proceeding WISDOM '13 Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, 2013

[9] Li, F., The Information Content of Forward-Looking Statements in Corporate Filings—A Naïve Bayesian Machine Learning Approach - *Journal of Accounting Research*, 2010, 48, (5), 1049-1102

[10] Pak, A., Paroubek, P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining - *In Proceedings of the Seventh Conference on International Language Resources and Evaluation*, 2010, 1320-1326

[11] SentiWordNet [WWW] http://sentiwordnet.isti.cnr.it/ (26.04.2015)

[12] About WordNet – WordNet [WWW] http://wordnet.princeton.edu/ (10.05.2015)

[13] CC BY-SA 3.0 - Creative Commons [WWW] http://creativecommons.org/licenses/by-sa/3.0/ (19.04.2015)

[14] AFINN [WWW] http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010 (26.04.2015)

[15] Open Database License (ODbL) v1.0 | Open Data Commons [WWW] http://opendatacommons.org/licenses/odbl/1.0/ (25.04.2015)

[16] The Stanford NLP (Natural Language Processing) Group [WWW] http://nlp.stanford.edu/software/ (15.03.2015)
[17] The GNU General Public License v3.0 - GNU Project - Free Software Foundation [WWW] http://www.gnu.org/licenses/gpl.html (15.03.2015)
[18] Apache OpenNLP - Welcome to Apache OpenNLP [WWW] https://opennlp.apache.org/ (15.03.2015)
[19] Apache License, Version 2.0 [WWW] http://www.apache.org/licenses/LICENSE-2.0 (15.03.2015)
[20] The OpeNER project [WWW] http://www.opener-project.eu/index.html (16.03.2015)
[21] Toutanova, K., Manning, C. D., Enriching the knowledge sources used in a maximum entropy part-of-speech tagger - *Proceeding EMNLP '00 Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics*, 2000, 13, 63-70
[22] Toutanova, K., Klein, D., Manning, C. D., Singer, Y., Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network - *Proceeding NAACL '03 Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 2003, 1, 173-180
[23] Finkel, J. R., Grenager, T., Manning C., Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling - *Proceeding ACL '05 Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 2005, 363-370
[24] Kea [WWW] http://www.nzdl.org/Kea/index.html (20.03.2015)
[25] maui-indexer - Maui - Multi-purpose automatic topic indexing - Google Project Hosting [WWW] https://code.google.com/p/maui-indexer/ (20.03.2015)
[26] InsideMaui - maui-indexer - General info on how Maui works and its components - Maui - Multi-purpose automatic topic indexing - Google Project Hosting [WWW] https://code.google.com/p/maui-indexer/wiki/InsideMaui (20.03.2015)
[27] Medelyan, O., Frank, E., Witten, I. H., Human-competitive tagging using automatic keyphrase extraction - *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 2009, 3, 1318-1327
[28] Business Insider [WWW] http://www.businessinsider.com/ (05.05.2015)
[29] Medelyan, O., Human-competitive automatic topic indexing : doktoritöö. University of Waikato, Hamilton, 2009
[30] BP profit hit by oil price fall - BBC News [WWW] http://www.bbc.com/news/business-32492438 (30.04.2015)
[31] Takahashi, S., Takahashi, M., Takahashi, H., Tsuda, K., Analysis of the Relation Between Stock Price Returns and Headline News Using Text Categorization - *KES 2007/ WIRN 2007, Part II, LNAI 4693*, 2007, 1339-1345

Lisa 1

Tabel 6 - Stanford Log-linear Part-Of-Speech Tagger märgendid

Märgend	Seletus
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PRP$	Possessive pronoun (prolog version PRP-S)
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun (prolog version WP-S)
WRB	Wh-adverb

Lisa 2

Tabel 7 - Positiivsete uudiste märksõnad

Uudise pealkiri	**Link**
Märksõnad	
What to Expect From Apple's Earnings Later Today	http://www.bloomberg.com/news/articles/2015-04-26/apple-iphone-sales-in-china-seen-surpassing-u-s-for-first-time
iPhone sales, New Year, Apple, demand, sales, U.S, iPhone, quarter, larger screen, customers	
Here's How Uber's Co-Founder Is Going to Take on Amazon and eBay	http://www.bloomberg.com/news/articles/2015-04-23/uber-s-co-founder-has-a-new-shopping-app-and-this-is-how-it-works
Garrett Camp, fortune helping, co founder, Camp, helping, people, Garrett, fortune, drivers, billion	
Cap Gemini to Buy Igate for $4 Billion to Take On Accenture	http://www.bloomberg.com/news/articles/2015-04-27/cap-gemini-to-buy-igate-for-about-4-billion-to-expand-in-u-s-
Cap Gemini, boost, U.S, billion, stock, billion euros, Cap, Igate, Gemini, Corp	
Fancy Photo App Startup VSCO Raises $30 Million More	http://www.bloomberg.com/news/articles/2015-04-24/fancy-photo-app-startup-vsco-raises-30-million-more
brand related, related hashtags, hashtag on Instagram, brand related hashtag, brand related hashtag on Instagram, related hashtags on Instagram, big, hashtag, venture capital, VSCO Cam	
Facebook Is Hiring Like Crazy	http://www.bloomberg.com/news/articles/2015-04-23/facebook-is-hiring-like-crazy
Zuckerberg, last year, Mark, Facebook, employees, sales, hiring, year, quarter, pace	
Facebook's mobile ad revenue jumps	http://money.cnn.com/2015/04/22/technology/social/facebook-earnings-first-quarter-2015/index.html?iid=SF_T_River
Facebook, mobile advertising, user engagement, company, social network, monthly active, active users, total monthly, advertising revenue, billion people	
Lenovo and Acer smartphones pack huge batteries	http://www.bbc.com/news/technology-32479717
Lenovo and Acer, batteries, Liquid X2, people, battery life, higher than, life, smartphones, battery, phone	
How Instagram is becoming a must-have for retailers	http://business.financialpost.com/investing/trading-desk/how-instagram-is-becoming-a-must-have-for-retailers?__lsa=b011-7bbc
social media, retailers, Facebook, Instagram, social, media, week, photo, Exstein, powerful	
As Samsung Galaxy S6 release date nears, positive reviews boost prospects for record shipments	http://business.financialpost.com/fp-tech-desk/personal-tech/as-samsung-galaxy-s6-release-date-nears-positive-reviews-boost-prospects-for-record-shipments
Galaxy S6, new Galaxy, record shipments, Samsung, Electronics, Galaxy, said, S6, demand, shipment	
Shopify's IPO success sets the stage for Ottawa to reclaim Silicon Valley North title	http://business.financialpost.com/entrepreneur/fp-startups/shopifys-ipo-success-sets-the-stage-for-ottawa-to-reclaim-silicon-valley-north-title?__lsa=1bf8-167f
industry, Silicon Valley, Valley North, Matthews, tech market, tech firms, investors, tech companies, based, Shopify	

Lisa 3

Tabel 8 - Negatiivsete uudiste märksõnad

Uudise pealkiri	**Link**
Märksõnad	
KFC 'dreadful,' says man who brought franchise to Britain	http://money.cnn.com/2015/04/14/news/kfc-uk-dreadful/index.html?iid=ob_article_footer&iid=obinsite
fried chicken, KFC franchise, fast food, fried, Kentucky, Allen said, company, chicken, Allen, said	
Google Looks to Buy Patents in Experiment to Lessen Lawsuits	http://www.bloomberg.com/news/articles/2015-04-27/google-looks-to-buy-patents-in-experiment-to-lessen-lawsuits
patent owners, owners, company, patents, Google, tech companies, licensing firms, licensing, looking, inventors	
Deutsche Bank shares fall on shake-up plan	http://www.bbc.com/news/business-32477137
Deutsche Bank, costs, business Postbank, sell, days, Bank, Deutsche, fallen, Shares, plans	
Ellen Pao faces $1m legal bill in sexism case	http://www.bbc.com/news/technology-32446358
Silicon Valley, Venture capital, Kleiner Perkins, Ms Pao, court, high profile, Pao, Silicon, lost, capital	
Deutsche Bank first-quarter profit falls by half as legal charges bite	http://www.reuters.com/article/2015/04/26/us-deutsche-bank-results-idUSKBN0NH0GI20150426
Deutsche Bank, investment banking, Deutsche, bank, billion euros, investment bank, net profit, million euros, percent rise, legal	
Ford recalling Fiesta, Fusion, Lincoln MKZ on door latch issue	http://www.reuters.com/article/2015/04/24/us-ford-motor-recall-idUSKBN0NF24J20150424
Lincoln MKZ, Ford, Lincoln, MKZ, Fiesta, Fusion, said, incidents in which an unlatched, unlatched door, incidents	
Baker Hughes plans new job cuts after quarterly loss on US$772M charge	http://business.financialpost.com/news/energy/baker-hughes-plans-new-job-cuts-after-quarterly-loss-on-us772m-charge?__lsa=b011-7bbc
Baker Hughes, earlier, billion, year earlier, year, Hughes, Baker, said it would cut, company, cent	
BlackBerry Ltd considers pulling out of Sweden and cutting 100 jobs in the process	http://business.financialpost.com/fp-tech-desk/blackberry-ltd-considers-pulling-out-of-sweden-and-cutting-100-jobs-in-the-process?__lsa=6b33-2d94
BlackBerry, offices in Sweden, offices, Sweden, considering, spokesperson, employees	
Les Pétroles Global Inc fined $1 million for fixing gas prices	http://business.financialpost.com/news/energy/les-petroles-global-inc-fined-1-million-for-fixing-gas-prices?__lsa=1bf8-167f
price, Les Pétroles, Pétroles Global, Quebec, Les Pétroles Global, Sherbrooke, Les, Pétroles, Global	
McDonald's Corp to close stores in Japan after food scandals slam sales	http://business.financialpost.com/news/retail-marketing/mcdonalds-corp-to-close-stores-in-japan-after-food-scandals-slam-sales
billion yen, McDonald, Japan, compared, yen, store, food, cent, loss, cuts	

Lisa 4

1. Chipotle has gone GMO-free. - http://money.cnn.com/2015/04/26/investing/chipotle-gmo-free/index.html?iid=HP_LN
2. KFC 'dreadful,' says man who brought franchise to Britain - http://money.cnn.com/2015/04/14/news/kfc-uk-dreadful/index.html?iid=ob_article_footer&iid=obinsite
3. Costco just killed my favorite credit card - http://money.cnn.com/2015/02/13/news/companies/costco-amex/index.html?iid=ob_article_footer&iid=obinsite
4. Siemens to axe 7,800 jobs - http://money.cnn.com/2015/02/06/news/companies/siemens-cuts-jobs/index.html?iid=ob_article_footer&iid=obinsite
5. Salesforce CEO promises women will get same pay as men - http://money.cnn.com/2015/04/24/technology/salesforce-equal-pay/index.html?iid=SF_T_River
6. Amazon lifts curtain on secretive $5 billion cloud business - http://money.cnn.com/2015/04/23/technology/amazon-earnings/index.html?iid=SF_T_River
7. Microsoft: The post-Windows company? - http://money.cnn.com/2015/04/23/technology/microsoft-earnings/index.html?iid=SF_T_River
8. Facebook's mobile ad revenue jumps - http://money.cnn.com/2015/04/22/technology/social/facebook-earnings-first-quarter-2015/index.html?iid=SF_T_River
9. Google launches 'Project Fi' wireless service - http://money.cnn.com/2015/04/22/technology/google-project-fi-wireless-service/index.html?iid=SF_T_River
10. The Apple Watch won't be available on April 24 - http://money.cnn.com/2015/04/16/technology/apple-watch-april-24/index.html?iid=SF_T_River
11. What to Expect From Apple's Earnings Later Today - http://www.bloomberg.com/news/articles/2015-04-26/apple-iphone-sales-in-china-seen-surpassing-u-s-for-first-time
12. Here’s How Uber’s Co-Founder Is Going to Take on Amazon and eBay - http://www.bloomberg.com/news/articles/2015-04-23/uber-s-co-founder-has-a-new-shopping-app-and-this-is-how-it-works
13. Cap Gemini to Buy Igate for $4 Billion to Take On Accenture - http://www.bloomberg.com/news/articles/2015-04-27/cap-gemini-to-buy-igate-for-about-4-billion-to-expand-in-u-s-
14. Fancy Photo App Startup VSCO Raises $30 Million More - http://www.bloomberg.com/news/articles/2015-04-24/fancy-photo-app-startup-vsco-raises-30-million-more
15. Tesla Motors Rises as Analysts Cite Energy Storage Opportunity - http://www.bloomberg.com/news/articles/2015-04-27/tesla-motors-rises-as-analysts-cite-energy-storage-opportunity

16. Google Looks to Buy Patents in Experiment to Lessen Lawsuits - http://www.bloomberg.com/news/articles/2015-04-27/google-looks-to-buy-patents-in-experiment-to-lessen-lawsuits
17. Deutsche Bank: This Week's Tesla Announcement Could Be a Bigger Deal Than Investors Realize - http://www.bloomberg.com/news/articles/2015-04-27/deutsche-bank-this-week-s-tesla-announcement-could-be-a-bigger-deal-than-investors-realize
18. Facebook Is Hiring Like Crazy - http://www.bloomberg.com/news/articles/2015-04-23/facebook-is-hiring-like-crazy
19. Nokia Targeting Apple, Alibaba and Amazon in Maps-Unit Sale - http://www.bloomberg.com/news/articles/2015-04-22/nokia-said-to-target-apple-alibaba-and-amazon-in-maps-unit-sale
20. Comcast Plans to Drop Time Warner Cable Deal - http://www.bloomberg.com/news/articles/2015-04-23/comcast-said-planning-to-withdraw-offer-for-time-warner-cable
21. BP profit hit by oil price fall - http://www.bbc.com/news/business-32492438
22. Will HSBC really quit the UK? - http://www.bbc.com/news/business-32446342
23. Apple to return more cash to investors as profits soar - http://www.bbc.com/news/business-32490926
24. Deutsche Bank shares fall on shake-up plan - http://www.bbc.com/news/business-32477137
25. EE revenue down as BT takeover looms - http://www.bbc.com/news/business-32477131
26. Whitbread profits up helped by rising Costa profits - http://www.bbc.com/news/business-32492608
27. Valve boss responds on game 'mod' row - http://www.bbc.com/news/technology-32480606
28. Lenovo and Acer smartphones pack huge batteries - http://www.bbc.com/news/technology-32479717
29. Jay Z says Tidal is 'doing just fine' with 770k subscribers - http://www.bbc.co.uk/newsbeat/article/32475661/jay-z-says-tidal-is-doing-just-fine-with-770k-subscribers
30. Ellen Pao faces $1m legal bill in sexism case - http://www.bbc.com/news/technology-32446358
31. Microsoft phones infringe patents: U.S. International Trade Commission judge - http://www.reuters.com/article/2015/04/27/us-microsoft-interdigital-decision-idUSKBN0NI23720150427
32. Investors question Deutsche Bank's overhaul - http://www.reuters.com/article/2015/04/27/us-deutschebank-restructuring-idUSKBN0NI0CW20150427
33. VW investors hope Piech exit may usher in change - http://www.reuters.com/article/2015/04/27/us-volkswagen-leadership-idUSKBN0NH0XQ20150427
34. Deutsche Bank first-quarter profit falls by half as legal charges bite - http://www.reuters.com/article/2015/04/26/us-deutsche-bank-results-idUSKBN0NH0GI20150426
35. Sony raises FY 2015 profit forecast to $2.5 billion: Nikkei - http://www.reuters.com/article/2015/04/25/us-sony-outlook-idUSKBN0NG04N20150425

36. Xerox cuts 2015 profit forecast due to strong dollar - http://www.reuters.com/article/2015/04/24/us-xerox-results-idUSKBN0NF12Q20150424
37. Ford recalling Fiesta, Fusion, Lincoln MKZ on door latch issue - http://www.reuters.com/article/2015/04/24/us-ford-motor-recall-idUSKBN0NF24J20150424
38. Credit Suisse CEO says post-crisis expansion was a mistake - http://www.reuters.com/article/2015/04/24/us-credit-suisse-shareholders-idUSKBN0NF0PL20150424
39. Amazon revenue beats, cloud computing more profitable than expected - http://www.reuters.com/article/2015/04/24/us-amazon-com-results-idUSKBN0NE2GR20150424
40. Microsoft profit, revenue beats Wall Street view; shares up - http://www.reuters.com/article/2015/04/24/us-microsoft-results-idUSKBN0NE2HE20150424
41. How Instagram is becoming a must-have for retailers - http://business.financialpost.com/investing/trading-desk/how-instagram-is-becoming-a-must-have-for-retailers?__lsa=b011-7bbc
42. Metro Inc profit jumps more than 15% despite hike in food costs - http://business.financialpost.com/news/retail-marketing/metro-inc-profit-jumps-more-than-15-on-higher-same-store-sales
43. Baker Hughes plans new job cuts after quarterly loss on US$772M charge - http://business.financialpost.com/news/energy/baker-hughes-plans-new-job-cuts-after-quarterly-loss-on-us772m-charge?__lsa=b011-7bbc
44. BlackBerry Ltd considers pulling out of Sweden and cutting 100 jobs in the process - http://business.financialpost.com/fp-tech-desk/blackberry-ltd-considers-pulling-out-of-sweden-and-cutting-100-jobs-in-the-process?__lsa=6b33-2d94
45. Jay Z's music streaming service Tidal flops, dropping out of iTunes' top 700 U.S. apps chart - http://business.financialpost.com/fp-tech-desk/personal-tech/jay-zs-music-streaming-service-tidal-flops-dropping-out-of-itunes-top-700-u-s-apps-chart?__lsa=64c7-c03c
46. As Samsung Galaxy S6 release date nears, positive reviews boost prospects for record shipments - http://business.financialpost.com/fp-tech-desk/personal-tech/as-samsung-galaxy-s6-release-date-nears-positive-reviews-boost-prospects-for-record-shipments
47. Shopify's IPO success sets the stage for Ottawa to reclaim Silicon Valley North title - http://business.financialpost.com/entrepreneur/fp-startups/shopifys-ipo-success-sets-the-stage-for-ottawa-to-reclaim-silicon-valley-north-title?__lsa=1bf8-167f
48. Les Pétroles Global Inc fined $1 million for fixing gas prices - http://business.financialpost.com/news/energy/les-petroles-global-inc-fined-1-million-for-fixing-gas-prices?__lsa=1bf8-167f
49. Barrick Gold Corp earnings miss estimates with output lower than expected - http://business.financialpost.com/news/mining/barrick-gold-corp-earnings-miss-estimates-with-output-lower-than-expected
50. McDonald's Corp to close stores in Japan after food scandals slam sales - http://business.financialpost.com/news/retail-marketing/mcdonalds-corp-to-close-stores-in-japan-after-food-scandals-slam-sales

Lisa 5

Tabel 9 - Kauguse arvutamise tulemused

Pealkiri	Inimene	Kaugus
What to Expect From Apple's Earnings Later Today	Pos	10,8185
EE revenue down as BT takeover looms	Pos	8,175257
Cap Gemini to Buy Igate for $4 Billion to Take On Accenture	Pos	7,908674
Fancy Photo App Startup VSCO Raises $30 Million More	Pos	7,75557
Here's How Uber's Co-Founder Is Going to Take on Amazon and eBay	Pos	6,134481
Shopify's IPO success sets the stage for Ottawa to reclaim Silicon Valley North title	Pos	5,643443
As Samsung Galaxy S6 release date nears, positive reviews boost prospects for record shipments	Pos	5,53362
Facebook Is Hiring Like Crazy	Pos	5,528787
BP profit hit by oil price fall	Neg	5,002943
Lenovo and Acer smartphones pack huge batteries	Pos	4,790761
Will HSBC really quit the UK?	Na	4,638871
How Instagram is becoming a must-have for retailers	Pos	4,570321
Facebook's mobile ad revenue jumps	Pos	4,527396
Apple to return more cash to investors as profits soar	Pos	3,533219
Amazon revenue beats, cloud computing more profitable than expected	Pos	3,033712
Costco just killed my favorite credit card	Na	2,796864
Tesla Motors Rises as Analysts Cite Energy Storage Opportunity	Pos	2,626823
Nokia Targeting Apple, Alibaba and Amazon in Maps-Unit Sale	Na	2,614879
VW investors hope Piech exit may usher in change	Pos	2,600984
Comcast Plans to Drop Time Warner Cable Deal	Na	2,550754
Whitbread profits up helped by rising Costa profits	Pos	2,28043
Sony raises FY 2015 profit forecast to $2.5 billion: Nikkei	Pos	2,213068
Credit Suisse CEO says post-crisis expansion was a mistake	Neg	2,135896
Valve boss responds on game 'mod' row	Na	1,934206
Microsoft profit, revenue beats Wall Street view; shares up	Pos	1,830393
Metro Inc profit jumps more than 15% despite hike in food costs	Pos	1,460409
Xerox cuts 2015 profit forecast due to strong dollar	Neg	1,001532
Salesforce CEO promises women will get same pay as men	Pos	0,753029
Amazon lifts curtain on secretive $5 billion cloud business	Pos	0,452987
Chipotle has gone GMO-free.	Pos	0,392869
Jay Z's music streaming service Tidal flops, dropping out of iTunes' top 700 U.S. apps chart	Neg	-0,39308
The Apple Watch won't be available on April 24	Neg	-0,7531
Microsoft phones infringe patents: U.S. International Trade Commission judge	Neg	-1,04866
Jay Z says Tidal is 'doing just fine' with 770k subscribers	Neg	-1,09905
Barrick Gold Corp earnings miss estimates with output lower than expected	Neg	-1,49461

Microsoft: The post-Windows company?	Pos	-1,69174
Investors question Deutsche Bank's overhaul	Neg	-1,92905
Deutsche Bank: This Week's Tesla Announcement Could Be a Bigger Deal Than Investors Realize	Pos	-2,06808
Siemens to axe 7,800 jobs	Neg	-2,63466
BlackBerry Ltd considers pulling out of Sweden and cutting 100 jobs in the process	Neg	-2,94329
Deutsche Bank first-quarter profit falls by half as legal charges bite	Neg	-3,07813
Ellen Pao faces $1m legal bill in sexism case	Neg	-3,13498
Google Looks to Buy Patents in Experiment to Lessen Lawsuits	Na	-3,56449
Google launches 'Project Fi' wireless service	Na	-4,38885
Baker Hughes plans new job cuts after quarterly loss on US$772M charge	Neg	-5,31858
KFC 'dreadful,' says man who brought franchise to Britain	Neg	-6,31137
Deutsche Bank shares fall on shake-up plan	Neg	-6,43119
Les Pétroles Global Inc fined $1 million for fixing gas prices	Neg	-8,65667
Ford recalling Fiesta, Fusion, Lincoln MKZ on door latch issue	Neg	-8,8856
McDonald’s Corp to close stores in Japan after food scandals slam sales	Neg	-9,86521

Lisa 6

Tabel 10 - Tonaalsuse analüüsi väljundid

Pealkiri	Inimene	SentiWordNet3 Summa	SentiWordNet3 Sõnu	AFINN-111 Summa	AFINN-111 Sõnu
Chipotle has gone GMO-free.	Pos	1,02456	193	14	21
KFC 'dreadful,' says man who brought franchise to Britain	Neg	-5,53642	237	4	15
Costco just killed my favorite credit card	Na	3,96635	211	12	22
Siemens to axe 7,800 jobs	Neg	-0,02431	84	-3	9
Salesforce CEO promises women will get same pay as men	Pos	2,30976	104	-3	13
Amazon lifts curtain on secretive $5 billion cloud business	Pos	1,53489	168	8	18
Microsoft: The post-Windows company?	Pos	-0,84043	136	9	12
Facebook's mobile ad revenue jumps	Pos	5,02817	218	18	18
Google launches 'Project Fi' wireless service	Na	-2,57191	180	2	20
The Apple Watch won't be available on April 24	Neg	1,57387	91	-5	7
What to Expect From Apple's Earnings Later Today	Pos	13,09003	409	55	49
Here’s How Uber’s Co-Founder Is Going to Take on Amazon and eBay	Pos	5,38423	318	36	18
Cap Gemini to Buy Igate for $4 Billion to Take On Accenture	Pos	8,67430	281	40	40
Fancy Photo App Startup VSCO Raises $30 Million More	Pos	6,83714	235	41	25
Tesla Motors Rises as Analysts Cite Energy Storage Opportunity	Pos	2,13179	83	8	12
Google Looks to Buy Patents in Experiment to Lessen Lawsuits	Na	-1,00871	93	0	10
Deutsche Bank: This Week's Tesla Announcement Could Be a Bigger Deal Than Investors Realize	Pos	-1,75906	154	16	17
Facebook Is Hiring Like Crazy	Pos	5,40446	167	14	13
Nokia Targeting Apple, Alibaba and Amazon in Maps-Unit Sale	Na	3,64990	155	6	17
Comcast Plans to Drop Time Warner Cable Deal	Na	2,82750	284	31	40
BP profit hit by oil price fall	Neg	4,19178	139	13	10
Will HSBC really quit the UK?	Na	6,87166	242	10	34
Apple to return more cash to investors as profits soar	Pos	3,82793	217	21	22
Deutsche Bank shares fall on shake-up plan	Neg	-3,16637	105	0	13
EE revenue down as BT takeover looms	Pos	5,14136	86	12	10

Whitbread profits up helped by rising Costa profits	Pos	2,20451	65	0	5
Valve boss responds on game 'mod' row	Na	2,78429	219	13	18
Lenovo and Acer smartphones pack huge batteries	Pos	5,99086	298	22	23
Jay Z says Tidal is 'doing just fine' with 770k subscribers	Neg	0,09591	94	4	8
Ellen Pao faces $1m legal bill in sexism case	Neg	-0,57839	123	-3	16
Microsoft phones infringe patents: U.S. International Trade Commission judge	Neg	1,02458	103	-4	19
Investors question Deutsche Bank's overhaul	Neg	0,04964	313	1	37
VW investors hope Piech exit may usher in change	Pos	3,94683	259	16	30
Deutsche Bank first-quarter profit falls by half as legal charges bite	Neg	-0,38034	232	-8	29
Sony raises FY 2015 profit forecast to $2.5 billion: Nikkei	Pos	1,84361	59	1	4
Xerox cuts 2015 profit forecast due to strong dollar	Neg	2,28421	151	4	14
Ford recalling Fiesta, Fusion, Lincoln MKZ on door latch issue	Neg	-1,98279	44	-6	6
Credit Suisse CEO says post-crisis expansion was a mistake	Neg	4,32654	176	-2	29
Amazon revenue beats, cloud computing more profitable than expected	Pos	5,30360	193	0	19
Microsoft profit, revenue beats Wall Street view; shares up	Pos	1,47453	103	9	8
How Instagram is becoming a must-have for retailers	Pos	2,54293	154	38	22
Metro Inc profit jumps more than 15% despite hike in food costs	Pos	0,75219	93	14	10
Baker Hughes plans new job cuts after quarterly loss on US$772M charge	Neg	-1,01730	52	-4	7
BlackBerry Ltd considers pulling out of Sweden and cutting 100 jobs in the process	Neg	0,75236	28	-3	1
Jay Z’s music streaming service Tidal flops, dropping out of iTunes’ top 700 U.S. apps chart	Neg	0,39764	145	9	14
As Samsung Galaxy S6 release date nears, positive reviews boost prospects for record shipments	Pos	5,47326	177	17	16
Shopify’s IPO success sets the stage for Ottawa to reclaim Silicon Valley North title	Pos	6,34935	389	44	43
Les Pétroles Global Inc fined $1 million for fixing gas prices	Neg	-2,17918	76	-20	13
Barrick Gold Corp earnings miss estimates with output lower than expected	Neg	0,08477	75	2	8
McDonald’s Corp to close stores in Japan after food scandals slam sales	Neg	-3,86746	99	-19	17