Tallinna Ülikool

Informaatika Instituut

NEUROVÕRKUDE KASUTAMINE FINANTSTURGUDE ANALÜÜSIKS

Bakalaureusetöö

Autor: Martin Ligema

Juhendaja: Erika Matsak

Autor: .....

Juhendaja: .....

Instituudi direktor: .....

Tallinn 2015

Autorideklaratsioon

Deklareerin, et käesolev bakalaureusetöö on minu töö tulemus ja seda ei ole kellegi teise poolt varem kaitsmisele esitatud. Kõik töö koostamisel kasutatud teiste autorite tööd, olulised seisukohad, kirjandusallikatest ja mujalt pärinevad andmed on viidatud.

.....

(kuupäev) (autor)

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina _____ _____ (sünnikuupäev:_____)

(autori nimi)

1. annan Tallinna Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose

_____

_____

_____

(lõputöö pealkiri)

mille juhendaja on _____,

(juhendaja nimi)

säilitamiseks ja üldsusele kättesaadavaks tegemiseks Tallinna Ülikooli Akadeemilise Raamatukogu repositooriumis.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tallinnas/Haapsalus/Rakveres/Helsingis, _____

(digitaalne) allkiri ja kuupäev

Sisukord

Sissejuhatus

Käesolev bakalaureusetöö esitab viisi kuidas algoritm õpib edukalt ennustama finantsturgude liikumisi, kasutades sisendina tekste sotsiaalmeediast, trende google otsingumootorist ja mineviku hinnainfot. Algoritmina kasutame kolmekihilist *feedforward* neurovõrku, mis klassifitseerib turusituatsioone vastavalt sisendile. Antud meetodit demonstreerime kasutades selleks Tesla Motors Company Inc. (TSLA) aktsiat.

Teema bakalaureusetööks sai valitud lähtuvalt huvist ja vajadusest masinõppe algoritmide vastu. Inimeste ja seadmete genereeritud andmete maht internetis kasvab. "*The total amount of information grew from 2.6 optimally compressed exabytes in 1986 to 15.8 in 1993, over 54.5 in 2000, and to 295 optimally compressed exabytes in 2007. This is equivalent to less than one 730-MB CDROM per person in 1986 (539 MB per person), roughly 4 CD-ROMper person of 1993, 12 CDROM per person in the year 2000, and almost 61 CD-ROM per person in 2007. Piling up the imagined 404 billion CD-ROM from 2007 would create a stack from the earth to the moon and a quarter of this distance beyond (with 1.2 mm thickness per CD).*" (Hilbert & López, 2011). Koos arvutite jõudluse kasvuga (Moore, 1998) võimaldab see, varem inimeste pärusmaaks peetud, otsustusprotsesse automatiseerida.

Börsifirmade tegevuse kohta on kättesaadaval palju tasuta informatsiooni ning esitatud töös üritame lahendada enda poolt väljatöötatud meetodi abil finantsinstrumendi ennustamise probleemi. Bakalaureuse töö eesmärk on vastata kolmele küsimusele, mis seonduvad turusituatsiooni klassifitseerimise probleemiga. Esiteks, uurida milline informatsioon aitab kaasa parema ennustuse tegemisele ? Teiseks, uurida kuidas mõjutab neurovõrgu teise kihi, ehk *hidden layer*i suurus resultaate. Ehk, kas hüpotees, rohkem neuroneid on parem, vastab tõele ? Kolmandaks, uurida kuidas klasside arvu muutmine mõjutab tulemusi.

Küsimustele vastuste leidmiseks jagame kogutud andmestikud päritolu järgi tükkideks. Koostame kolm erinevat neurovõrgu sihtmärk klassi. Valime kolm neurovõrgu arhitektuuri ning treenime need kogutud andmetega, võttes arvesse kõiki võimalusi ning võrdleme tulemusi

Bakalaureusetöö käigus demonstreerime ka sotsiaalmeedia sõnumitest hinnangu eraldamise tööriistu. Tõlgendame TSLA firma CEO Elon Musk Twitteri voogu läbi AlchemiAPI (application programming interface). Võtame kasutusele *retweet*imise, *favourite*mise andmeid, Elon Muski teise börsifirma Solar City ltd aktsia hinna, Tesla Motors Company ltd

aktsia hinna ajaloo ning Teslaga seotud Google otsingute trende. Antud uurimustöö lisaeesmärk on uurida, kas neurovõrguga treenitud klassifitseerija suudab kõiki neid erinevaid näitajad arvesse võttes, ennustuste täpsuses ületada neid uurimusarendusi, milles on arvestatud ainult sotsiaalmeediast pärit infoga.

Ülevaade olemasolevast

Tallinna Ülikoolis on varem kaitstud bakalaureusetöö, mis baseerus tehnilise analüüsi teooriatel (Lass & Kippar, 2014). Mis puutub aktsiate hinna lühiajalisse ennustamisesse kasutades tehnilise analüüsi meetodeid, toetudes da Costa, Nazário, Bergo, Sobreiro, & Kimura 2015 aastal Brasiilias tehtud uurimuse tulemustele, usume, et on paremaid viise kuidas hinda ennustada: „The results indicate that while the studied techniques lead to a high probability of obtaining a return that exceeds the investment value, they have little power of predictability in the Brazilian market. In relation to the passive buy strategy, only the smallest part of the obtained returns outweighs the results of the buy-and-hold strategy.“ (da Costa, Nazário, Bergo, Sobreiro, & Kimura, 2015).

Tehnilise analüüsi käigus uuritakse turul olevat nõudlust ja pakkumist ning proovitakse leida trende, mis jätkuvad tuleviks. „Technical analysis studies supply and demand in a market in an attempt to determine what direction, or trend, will continue in the future. In other words, technical analysis attempts to understand the emotions in the market by studying the market itself, as opposed to its components.“ (Janssen, Langager, & Murphy, n.d.) Enamik tehnilise analüüsi rakendusi, mis töötavad nendes valdkondades tuginevad kaupleja poolt käsitsi tehtud funktsioonidel. On selge, et kõnealuste süsteemide toimimine sõltub suuresti funktsiooni koostaja professionaalsest kompetensist, kvaliteedist ja oskusest turu olukordi ette näha. Lähtuvalt efektiivse turu hüpoteesist (Fama, 1965)(Veskimägi, 2006) liigutavad aktsiaid turgudel enamasti uue informatiooni, ehk uudiste tekkimise tagajärjel. Uudised on iseenesest ettearvamatud seega, kui eesmärgiks on turu liikumisi teistest kiiremini tajuda, on mõistlik õpetada arvuti uudiseid lugema. Masinõppe vaatenurgast on turusituatsioonide hindamine klassifitseerimise probleem. Logistilisel regressioonil baseeruvad mitmekihilised *feedforward* neurovõrgud on tõestanud ennast kui üks lihtsamaid ning tõhusamaid viise andmete klassifitseerimiseks (LeCun & Bengio, 1995).

Zhang, Fuehres ja Gloor uurisid oma 2011 aasta uurimuses "Predicting Stock Market Indicators Through Twitter "I hope it is not as bad as I fear" ", kas twitteri postitused ennustavad aktsiaturgude indekseid nagu Dow Jones, NASDAQ ja S&P 500. Nad kasutasid bag-of-words andmete kogumise meetodit, ning leidsid, et säutsud Twitteris korreleeruvad negatiivselt nende indeksitega. Nad järeldasid, et kui inimesed on kollektiivselt emotsionaalsed, ehk säutsutakse emotsionaalseid, murelikke sõnumeid, siis indeksid liiguvad alla. Kui inimestel on vähem hirme ja muresid ja siis indeksid liiguvad üles (Zhang, Fuehres, & Gloor, 2011). Kui eesmärgiks on uudiste ja hinnainfo vahelist seost arvutile tõlgendada siis kõige lihtsam viis selleks oleks uurida sõnade sageduse ja turutrendide vahelist statistilist seost, nagu eelmainitud uurimustöös tehti. Kuigi see tehnika võib olla tõhus, kaotatakse ära sõnumi kontekstiga kaasnev arvamuse ehk hinnang. Käesolevas töös, uurime lähemalt Tesla CEO Elon Muski säutsude seost Tesla aktsiaga (TSLA). Ühe indiviidi säutsude koguse juures peame igat tweeti käsitlema eraldi ning kasutamse selleks AlchemyAPId. Keskendume hinnangule ja säutsu emotsioonidele ning uurime kas see võiks ennustada firma aktsia liikumisi. Tweetidele hinnangu andmiseks kasutame IBM korporatsiooni tütarettevõtte AlchemyAPI Sentiment Analysis APIt (application programming interface).

Antud bakalaureusetöös ei uurita täpselt milline on uudiste ja finantsturgude liikumise vaheline seos. Toetume varasemalt kinnitust saanud faktidele, et seos on olemas. (Alanyali, Moat, & Preis, 2013), (Q. Li et al., 2014), (Smailović, Grčar, Lavrač, & Žnidaršič, 2014), „We find that, when information can be identified and that the tone (i.e., positive versus negative) of this information can be determined, there is a much closer link between stock prices and information.“ (Boudoukh, Feldman, Kogan, Richardson, & Roll, 1988). Twitteri ja aktsiaturgude vahelist korrelatsiooni on varemgi uuritud, ning on kinnitatud selle olemasolu. (Zhang et al., 2011). Tesla Motors Company aktsia hinna ja firma CEO Elon Muski tweetide vahelist seosest on isegi meedias juttu olnud (Assis, 2015).

Bollen, Mao ja Zeng uurisid 2011 aastal Twitteri säutsude ja Dow Jones vahelist seost. Kasutades OpinionFinder süsteemi (http://mpqa.cs.pitt.edu/opinionfinder/) , mis eraldab igast lausest (tuju) eraldi. Leiti, et nende meetoditega suudetakse 86,7% täpsusega ennustada Dow Jones aktsia indeksi liikumisi (Bollen, Mao, & Zeng, 2011). See saab olema antud bakalaureuse töös eesmärk mida üritame ületada. Uurime, kas Twitteri andmetele ka muude andmete lisamine aitaks ennustamist täpsemaks teha.

Jasmine Smailovic, Mihar Cracar, Nada Lavarc ja Martin Znidarsic uurisid 2013 aasta uurimuses „*Stream-based active learning for sentiment analasys in the financial domain*“

kuidas twiteri mõtteavalduste tuju võimaldab finantsinstrumente ennustada. Leitakse, et kui järgida konkreetset aktsiat, muutused muutuseid säutsude emotsionaalses toonis võivad olla edukad indikaatorid päeva sulgemis hinnas. (Smailović et al., 2014)

Uurimustöös „ *News impact on stock price return via sentiment analasys*“ (X. Li, Xie, Chen, Wang, & Deng, 2014), leitakse, et Hong Kongi Stock Exchange aktsia turul individuaalse uudises sisalduva sentimendi mõõtmisel saavutatakse parmaid ennustavaid tulemusi kui bad-of-words meetodiga.

Varasemate tööde baasil võime järeldada, et Twitteri säutsude najal on võimalik ennustada aktsiate liikumisi. Järedame ka, et käsitledes igat säutsu eraldi võime saavutada paremaid tulemusi kui *bag-of-words* meetodi kasutades.

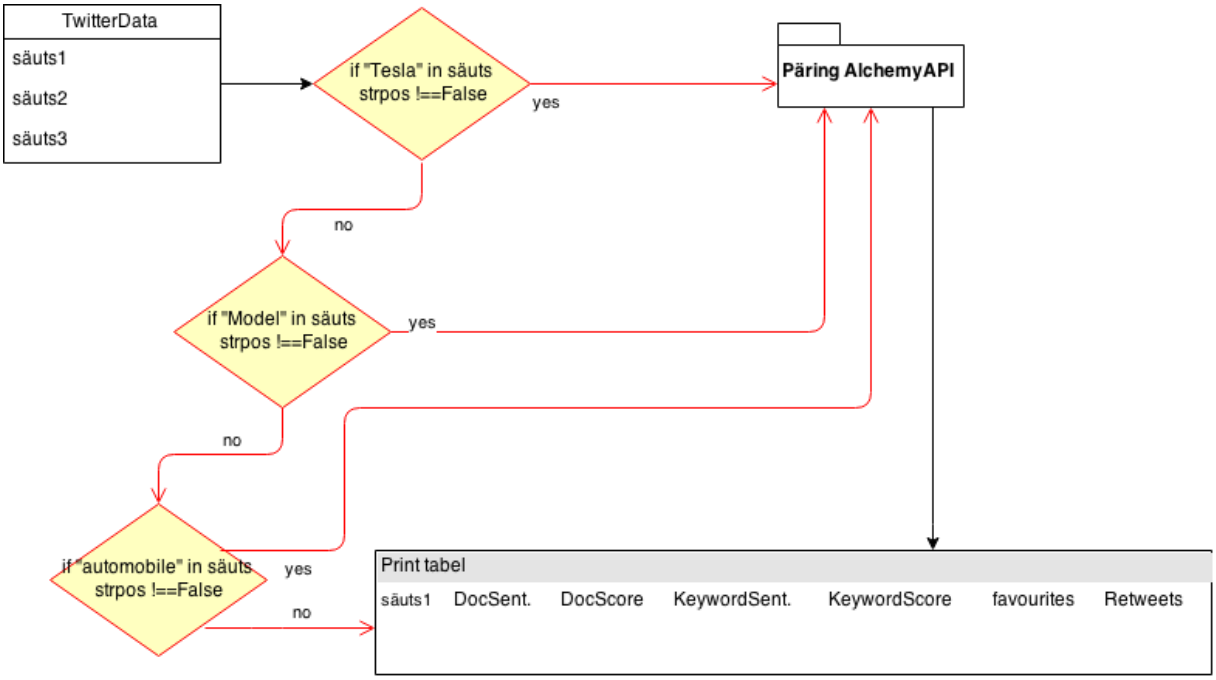
Eelnevalt mainitud uurimustööd ei käsitlue eraldi milllise informatsiooniga saaks, paremaid ennustusi teha: säutsudega, mineviku hinnainfoga, Google otsingumootori trendide baasil. Eelnevate uurimuste käigus ei kasutatud tehsilikke neurovõrke, seega järgnevalt uurime, kui kasutada lihtsat kolmekihilist neurovõrku, kas neuronite arv on oluline määraja paremate tulemuste saamiseks. Lisaks eelmainutle , ei pööratud teistes uurimustöödes tähelepanu Twitteri retweetimiste ja favouriteäimiste võimekusele aktsiate homset hinda ennustada, antud arendusuurimuse käigus vaatame kas seal võib olla seost.

Andmete Kogumine

Eelnevatest uurimustest lähtuvalt on tõendeid , et Twitteris säutsud võivad edukalt ennustada homseid aksita hindu. Andmete kogumise käigus kogun firma CEO Twitter voo, firma CEO säutsudega kaasnevad andmed (favourites, retweets), Google trends „Tesla“ otsingu termingia kõige enim korreleeruvat otingud, ajaloolise hinnainfo, Solar city aktsiahinna ajaloo ja Tesla aktsiahinna ajaloo. Esimeseks andmete kogumise etapiks on koguda twitterist kõik Elon Muski tehtud säutsud. Loodud rakenduse aluseks on võetud Tallinna Ülikooli Informaatika instituudi aine "IFI6093 Veebirakenduste kasutajaliidesed" üks tunnitöödest. Antud aine käigus omandatud teadmised olid käesoleva prototüübi loomise käigus palju abiks. Twitteris on võimalik säutse uuesti säutsuda, ehk retweetida või favourite'ida, ning lisame need ka oma andme tabelitesse. Twittersi API kaudu hangime andmed Tesla CEO Elon Muski kontolt alates 04.06.2010 kuni 15.04.2015. Kokku on selle perioodiga tehtud 864 individuaalset tweeti. Twitterist laeme andmetena alla tweedi kuupäeva, sisu, *retweet*imiste ja

*favourite*mise arvu. Twitter API väljastab andmeid säutsude kaupa, kuid käesolevas uurimuses soovime ennustada homset aktsia hind, seega tuleb säutsud kronoloogilisse järjekorda seada, jättes tühikud päevadele, kus säutse polnud. Selleks koostasime programmi andmetesorteerimine.php millega saab tutvuda (Lisa1).

Peale kronoloogilisse järjekorda seadmist kasutame programmi nimega GetSentiment.php, mis loeb säutse sisaldavast failist säutsud. Uurib kas see sisaldab sõnu „Tesla“, „Model“ või „automobile“, ning kui sisaldab siis edastatakse see sõnum Alchemy APIsse. Kui säuts ei sisalda neid sõnu siis priditakse välja , et sõnum ei sisaldanud võtmesõnu ning liigutakse edasi järgmise säutsu juurde. Võtmesõnu sisaldavate säutsude sorteerimiseks teistest, kasutame PHP keele funktsiooni strpos (“PHP: strpos - Manual,” n.d.). Funktsioon strpos väljastab numbri, mitmes string on otsitud võtmesõna stringi, antud juhul säutsu, sees. Kuna meie eesmärgiks on teada saada, kas üldse võtmesõna peitub säutsu sees, siis kasutame loogilist tehet „!==false“, ehk tõeseks vastuseks loeme kõik vastused, mis rahuldavad strpos funktsiooni tingumusi.

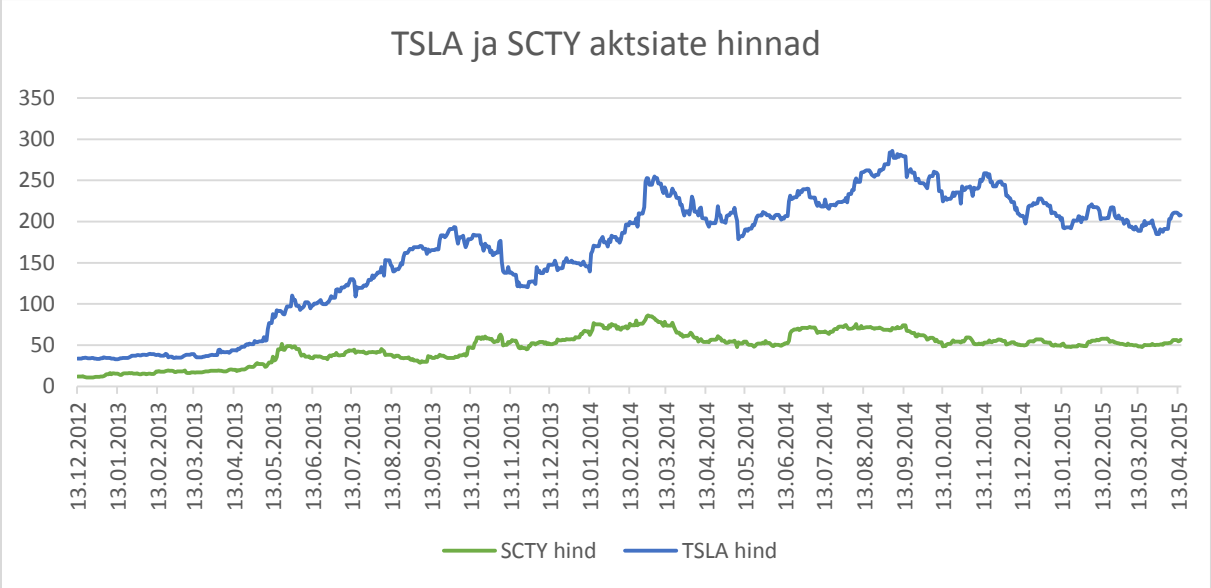


Joonis 1 programm GetSentiment.php.

Solar City ltd. (SCTY) on ettevõte mille üks asutajates on samuti Elon Musk. Kuigi puuduva uuringud, mis tõestaks SCTY ja TSLA aktsia vahelist seost on sellest võimalikust seosest artikkel kirjutatud (Duggan, 2015). Kaasame SCTY ajaloolise hinnainfo andmetesse. Info hangime samuti Google Spreadsheets rakendusest.

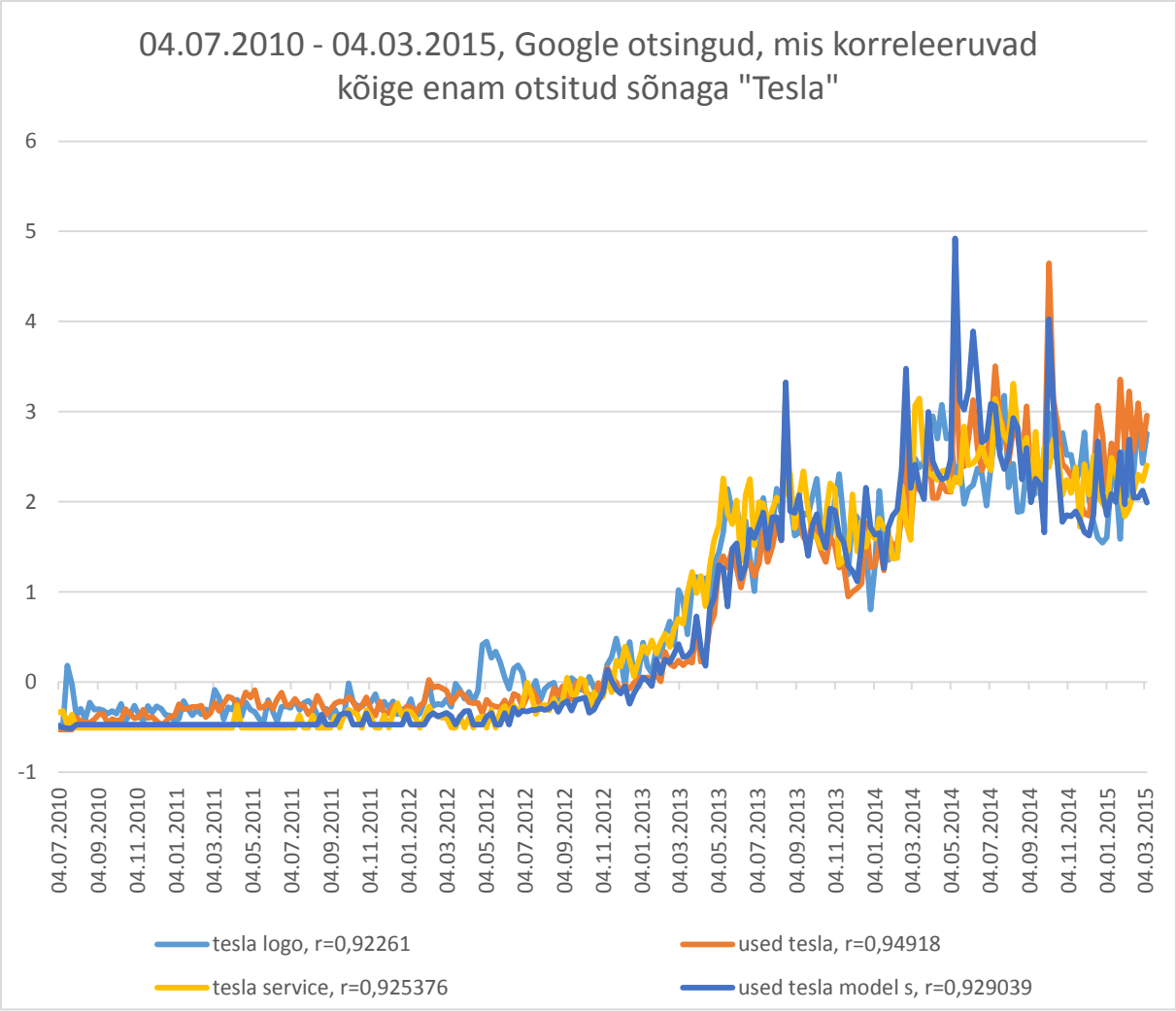
Tesla ja Solar city aktsiate hinnainfo saab kätte google spreadheetsilt. Tesla Motors Company aktsia (TSLA) hinnainfot saab tasuta küsida Google Spreadheets veebirakendusest funktsiooniga „=googlefinance(„aktsia nimi“, „hinna mõõtlmise aeg: open/high/low/close“, „algus kuupäev“, „lõpp kuupäev“, „ periooni intervall: daily/weekly/monthly“)“. Päeva jooksul toimunud uudiste mõju saame hinnata päeva lõpus, seega hinna mõõtmise ajaks tuleb valida kauplemise päeva lõpp, ehk close. Perioodi minimaalne intervall on päevane, ehk daily. Aktsia hinnaliikumise hangime alates 03.06.2010 kuni 15.04.2015. Kusjuures Solar City aktsia hinna liikumise lineaarne korrelatsioonikordaja ehk Pearsoni korrelatsioonikordaja r = 0.86 , ehk on tugev korrelatsioon (*Joonis* 3.).

Joonis 2 Tesla aktsia hind alates 29.06.2010 kuni 28.02.2015 Allikas: google.finance.com.



Joonis 3 Tesla ja Solar City aktsia hinnad alates 13.12.2012 kuni 13.04.2015. Allikas: google.finance.com

Google otsingumootori trendidest andmeid saab hankida leheküljel https://www.google.com/trends/correlate/ . Google Correlate arvutab Pearson korrelatsioonikordaja iga nädala tagant tehtud otsingute sageduse ja otsitud termini otsingute sageduse vahel. Laeme alla sada kõige suurima korrelatsioonikordajaga otsingutermini otsimissagedused ajavahemikus 04.01.2004 kuni 08.03.2015. Allikas: (https://www.google.com/trends/correlate/csv?e=tesla&t=weekly&p=us). Googe trends väljastab 100 kõige enim Pearsoni järgi korreleeruvat otsinguterminti. Kuigi enamik neist on seotud Tesla iseendaga leidub ka huvitavamaid näiteks „*power in the name of Jesus*“ korrelerub tesla otsingutega viieaastase perioodi juuksul Pearsoni koefitsiendi järgi 0,752.



Joonis 4. Valik otsinguterminitest, mis korreleerusid otsinguga "Tesla" Allikas google.trends.com.

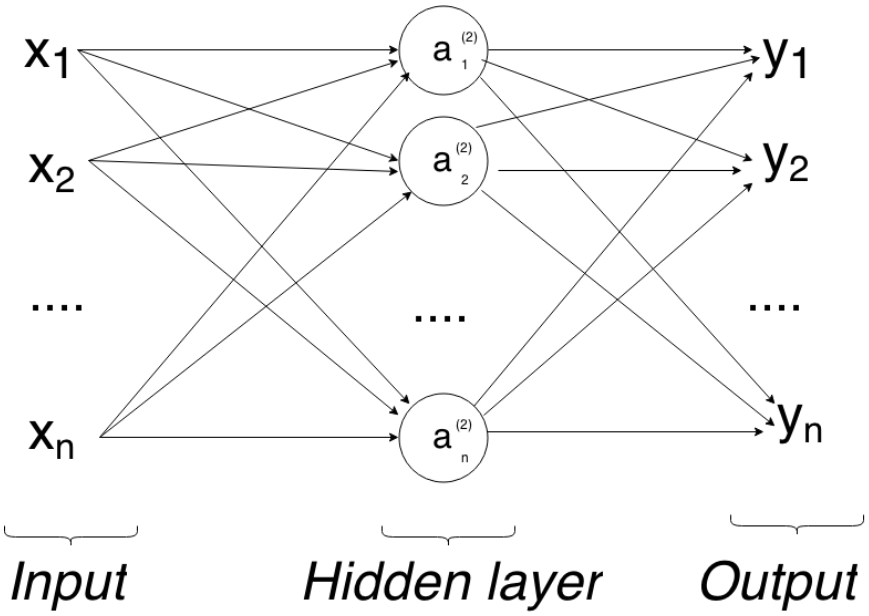
Andmegruppide moodustamine

Bakalaureuse töö üks eesmärke on võrrelda millise sisend informatsiooni baasil oskab neurovõrk kõige täpsemaid ennustusi teha. Selleks peame kogutud andmed jagama gruppidesse. Ütleme, et $\mathbf{X_1}$ on twitterist pärist säutsude maatriks kujuga 1232 x 5 , $\mathbf{X_2}$ on Google trends otsingumootorist pärit korrelatsioonide tabel kujuga 1232 x 100. $\mathbf{X_3}$ on Tesla ja Solar City mineviku hinnainfo maatriks kujuga 1232 x 41. $\mathbf{X_4}$ on kõik eelnevad andmed kokku panduna maatiksisse kujuga 1232 x 146. Sihtmärk, ehk *target* vektoriks saab olema Tesla Motors Company päeva lõpu hind. Selleks, et uurida kuidas kuidas klasside arvu suurus mõjutab ennustuste täpsust jagame hinnainfo kolme erinevasse klassi. $\mathbf{Y_1}$ saab olema 9 klassiga sihtmärk, ehk ühe klassi täpsus seljuhul on +/- 15,01 dollarit. $\mathbf{Y_2}$ saab olema 19

klassiga sihtmärk, ehk ühe klassi täpsus seljuhul on +/- 7,11 dollarit ning $\mathbf{Y_3}$ olgu 33 klassiga sihtmärk, ehk ühe klassi täpsus seljuhul on +/- 4,09 dollarit.

Neurovõrgud

Feedforward neurovõrgud on algoritmid ,mida saab treenida klassifitseerima keerulisi andmeid. Treenimise faasis kasutatakse sihtandmeid (target classes) selleks, et logistilise regressiooni algoritmid ehk neuronid, omistaks soovitud väärtused (Ng, 2010) (Anderson, 1996) . Käesoleva uurimustöö käigus kasutame kolmekihilsti neurovõrku, mis koosneb info sisestamise kihist ehk input layerist, peidetud kihist ehk *hidden layeri*st ja väljundikihist ehk output layerist (Joonis 5). Neurovõrgu esimese, input layeri ja viimase, ehk output layeri suurused on kindlalt defineeritud, vastavalt ülesandele mida lahendada soovitakse. See tähendab, et kui on 1232 erineat sisendit siis neurovõrgu sisendkiht **X** suurus on 1232. Neurovõrgu output layeri suurus on samuti ettemääratud, nimelt klasside arvuga mida soovitakse kasutada. Antud juhul on selleks, eelnevalt määratletud, kolm klassi $\mathbf{y_1 = 9, y_2 = 19, y_3 = 33.}$ Sisendkihi **X** eesmärk on edastada informatsiooni neurovõrgu teisele kihile $\mathbf{a^{(2)}}$. Neurovõrgu teise kihi, ehk *hidden layer*i $\mathbf{a^{(2)}}$ suurust saab määrata. Uurime kuidas neurovõrgu teise kihi suurus mõjutab klassifitseerimise võimekust. Katsetame kolme erinevat neurovõrgu suurust teisel kihil : $\mathbf{a_5^{(2)}}$ = 5 neuronit, $\mathbf{a_{25}^{(2)}}$ = 25 neuronit ja $\mathbf{a_{150}^{(2)}}$ = 150 neuronit. Kus $\mathbf{a^{(2)}}$ on neurovõrgu teise kihi tähis ning $\mathbf{a_n^{(2)}}$ kus **n** tähistab neuronite arvu mida näite käigus kasutatakse. Prototüüpide tegemiseks kasutame programmi Matlab (MATLAB, 2013). 70% andmetest kasutame, et treenida neurovõrku (*training set*), 15% andmetest kasutasime tulemuste kontrollimseks (*test set*) ja ülejäänud 15% kasutasime neurovõrgu ristvalideerimiseks (*cross-validation*).



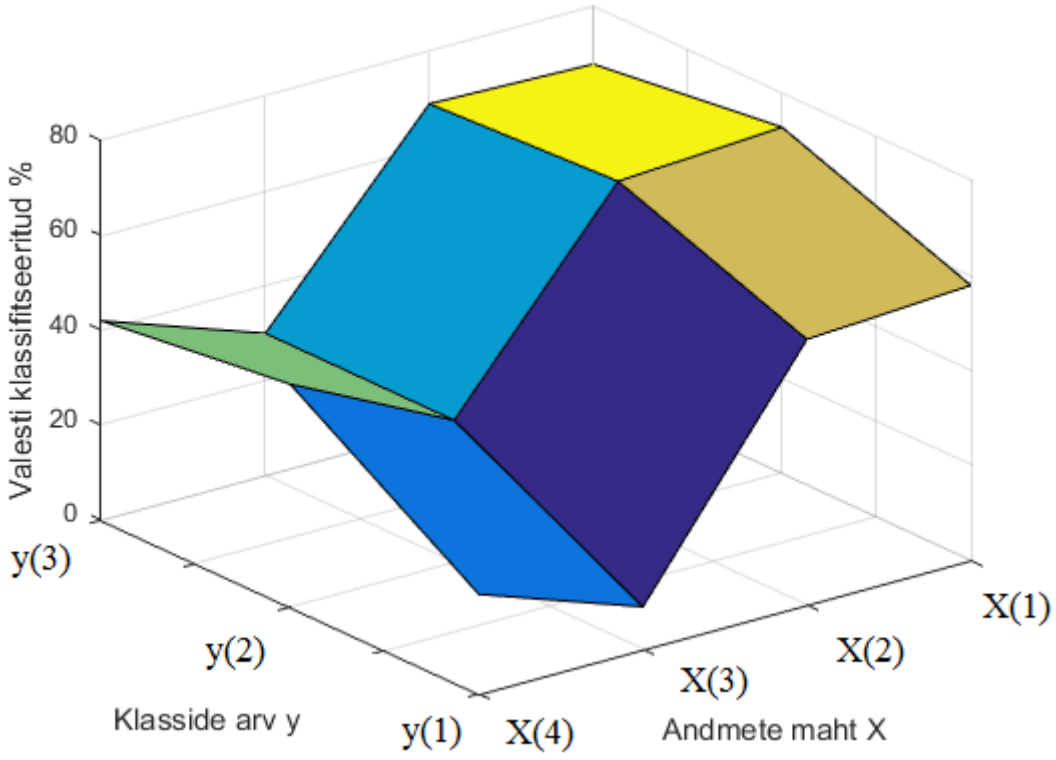
Joonis 5. Lihtsustatud neurovõrgu skeem. Kus ***n*** *tähistab neurovõrgu teise kihi* $\boldsymbol{a}^{(2)}$ *neuronite arvu.*

Tulemused

Antud tulemuste saamiseks on mõõdetud, mitu protsenti treening andmetest klassifitseeriti valesti vastavate sisendandmetega **X** ja klasside arvuga **y**. Kuvame saadud tulemused järgnevates tabelites Tabel 1, Tabel 2, Tabel 3 ja joonistes Joonis 6, Joonis 7 ja Joonis 8. Tulemuste tõlgendus on toodud järgmises alampeatükis (vt. Tulemuste tõlgendus).

Tabel 1. Kasutades 5 neuronit *hiddel layer*is.

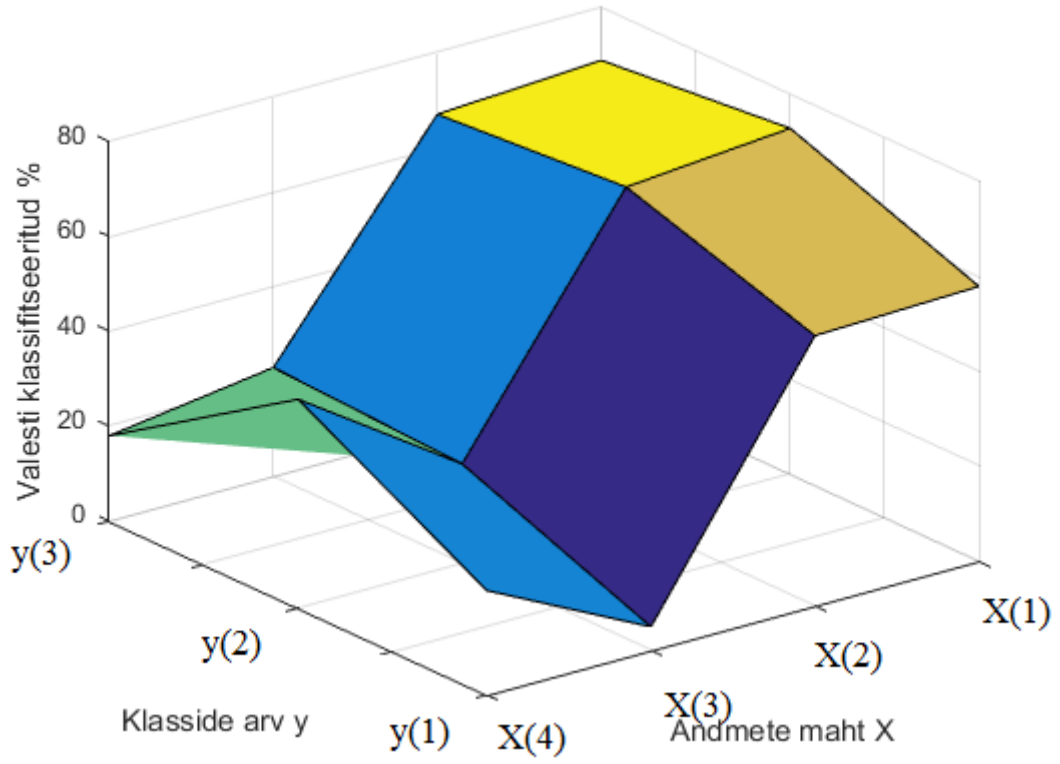
Kasutades $\mathbf{a_5^{(2)}}$ = 5 neuronit		Andmete maht			
		$\mathbf{X_4}$	$\mathbf{X_3}$	$\mathbf{X_2}$	$\mathbf{X_1}$
Klasside arv	$\mathbf{y_1}$	21 %	9 %	56 %	58 %
	$\mathbf{y_2}$	47 %	30 %	71 %	73 %
	$\mathbf{y_3}$	42 %	30 %	69 %	68 %



Joonis 6. Tulemused kasutades 5 neuronit teises neurovõrgu kihis.

Tabel 2. Tulemused 25 neuroniga neurovõrgu teises kihis

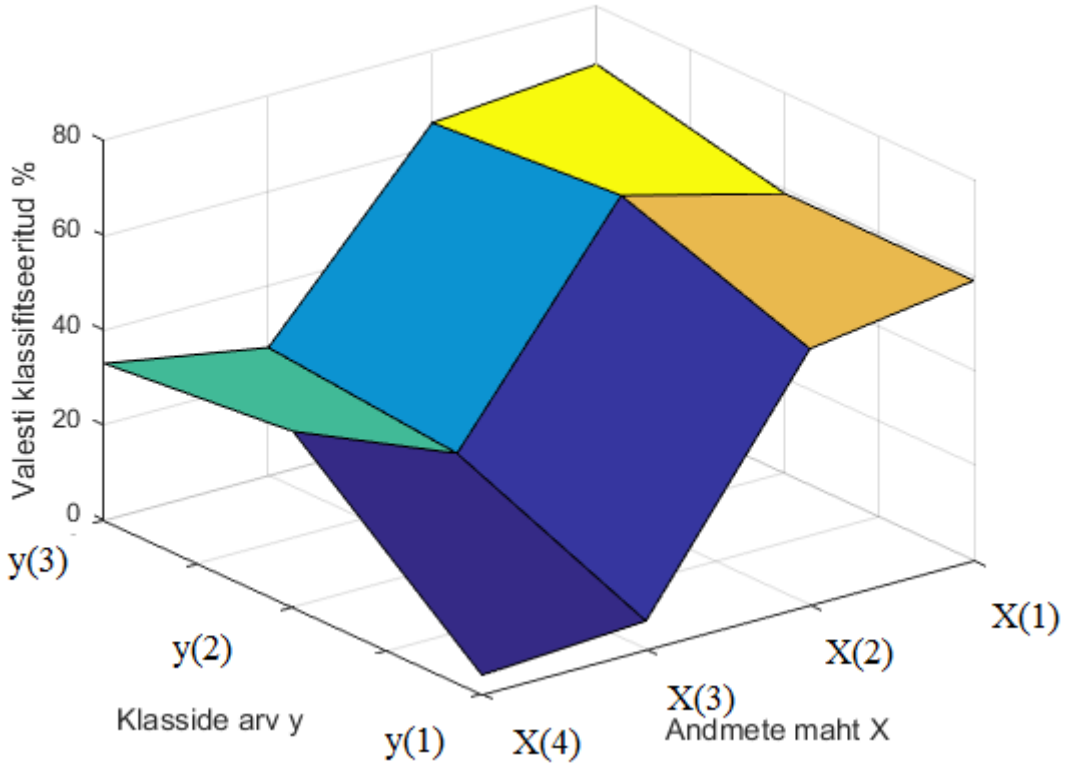
Kasutades $a_{25}^{(2)}$ = 25 neuronit		Andmete maht			
		X_4	X_3	X_2	X_1
		Kokku	**Hinnainfo**	**Google**	**Twitter**
Klasside arv	y_1	22 %	5 %	57 %	58 %
	y_2	44 %	21 %	70 %	73 %
	y_3	**18 %**	23 %	67 %	69 %



Joonis 7 . Tulemused kasutades 25 neuronit teisel kihil.

Tabel 3. Tulemused 150 neuronit neurovõrgu teises kihis

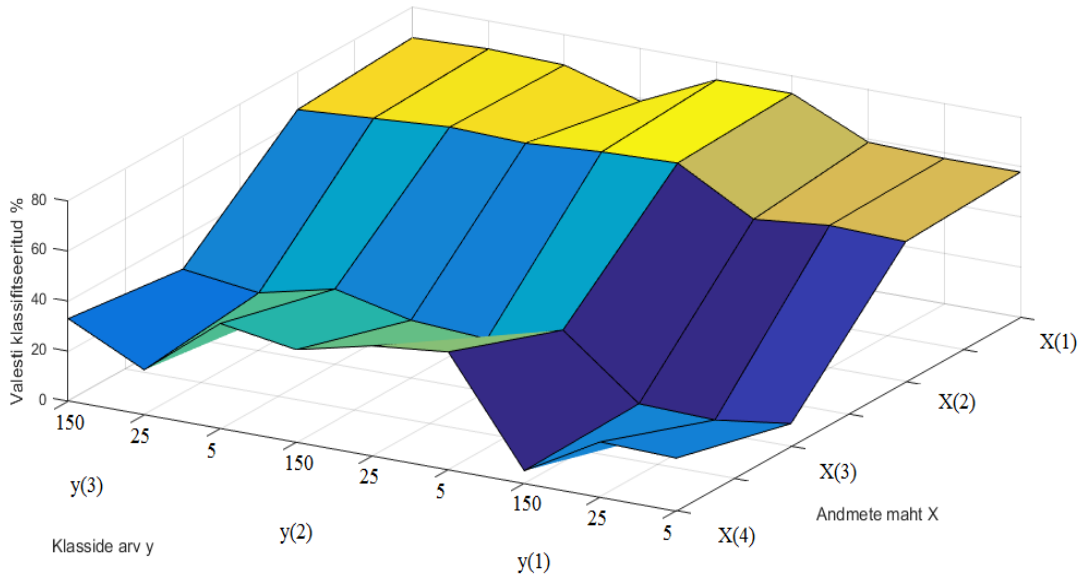
Kasutades $a_{150}^{(2)}$ = 150 neuronit		Andmete maht			
		X_4	X_3	X_2	X_1
		Kokku	**Hinnainfo**	**Google**	**Twitter**
Klasside arv	**y_1=9**	**4 %**	6 %	54 %	59 %
	y_2=19	37 %	23 %	68 %	59 %
	y_3=33	33 %	27 %	65 %	68 %



Joonis 8. Tulemused kasutades 150 neuronit neurovõrgu teisel kihil.

Tabel 4. Koondtabel

			Andmete maht					
			Kokku	**Hinnainfo**	**Trends**	**Twitter**		
			X_4	**X_3**	**X_2**	**X_1**	Keskmine viga $a_n^{(2)}$ kohta	keskmine viga **y**
Klasside arv	**y_1=9**	**n = 5**	21 %	9 %	56 %	58 %	36 %	34 %
		n = 25	22 %	5 %	57 %	58 %	36 %	
		n = 150	**4 %**	6 %	54 %	59 %	31 %	
	y_2=19	**n = 5**	47 %	30 %	71 %	73 %	55 %	51 %
		n = 25	44 %	21 %	70 %	73 %	52 %	
		n = 150	37 %	23 %	68 %	59 %	47 %	
	y_3=33	**n = 5**	42 %	30 %	69 %	68 %	52 %	48 %
		n = 25	**18 %**	23 %	67 %	69 %	44 %	
		n = 150	33 %	27 %	65 %	68 %	48 %	
		Keskmine viga **X** järgi	30 %	19 %	64 %	65 %		



Joonis 9. Koondtabel.

Tulemuste tõlgendus

Käesoleva bakalaureusetöö eesmärkideks oli leida vastused kolmele küsimusele. Esimeseks uurimiseesmärgiks oli selgitada milline informatsioon aitab kaasa parema ennustuse tegemisele. Üleüldine trend ,läbi kõigi andmemahtude, näitab, et kõige paremini aitab homset hinda ennustada andmete grupp $\mathbf{X_3}$ ja $\mathbf{X_4}$. Grupp $\mathbf{X_3}$, aktsiahindadegrupp, klassifitseeris valesti keskmiselt 19 % turuolukordi. Grupp $\mathbf{X_4}$, kus kõik andmed olid kokku pandud klassifitseeris valesti keskmiselt 30 % turuolukordi. Oluline esile tuua andmegrup $\mathbf{X_4}$ ja klassid y3 katset kui neuronite arv a = 25, kus valesti klassifitseeriti ainult 18 % näiteid (Joonis 9.). Meenutame siinkohal, et y3 grupis oli 33 klassi, mis andis ennustuse täpsuseks +/- 4,09 dollarit. Võimekus ennustada homset aktsiahinda ligikaugu 8 dollarise kõikumise sees 82 % tõenäosuse juures märkimisväärne saavutus.

Veel on oluline esile tuua kõige kõrgema ennustuse täpsusega katse tulemust. See oli andmegrupi X4 ja klasside y1 puhul, kui kasutati 150 neuronit. Kasutades kõige suuremat andmegruppi, kõige väiksemat klasside arvu ja kõige suuremat neuronite hulka saavutati kõige usaldusväärsem klassifitseerimise tulemus 96 %. See tähendab, et suudame ennustada homset aktsiahinna liikumist vahemikus +/- 15,01 dollarit 96 protsendilise tõenäosusega.

Ainult Twitteri andmete baasil tehtud ennustuste vigade arv on suur, 65 % ennustusi läksid valesti. Kuid lisades twitteri andmed kokku teiste andmetega siis vigade arv väheneb 65 protsendilt 30 protsendile. Põhjuseid ,miks twitteri baasil tehtud ennustused nii kehvad on, võib olla mitmeid, kui toon välja siinkohal fakti, et kasutada oli umbes 800 erinevat säutus, millest veel osad ei olnud üldse Tesla Motors Companu teemalised. Võrreldes näiteks (bollen, viide) kasutasin oma uuringus kokku üle 9 miljoni säutsu ning saavutasid sealjuures 86,7 % lise täpsuse.

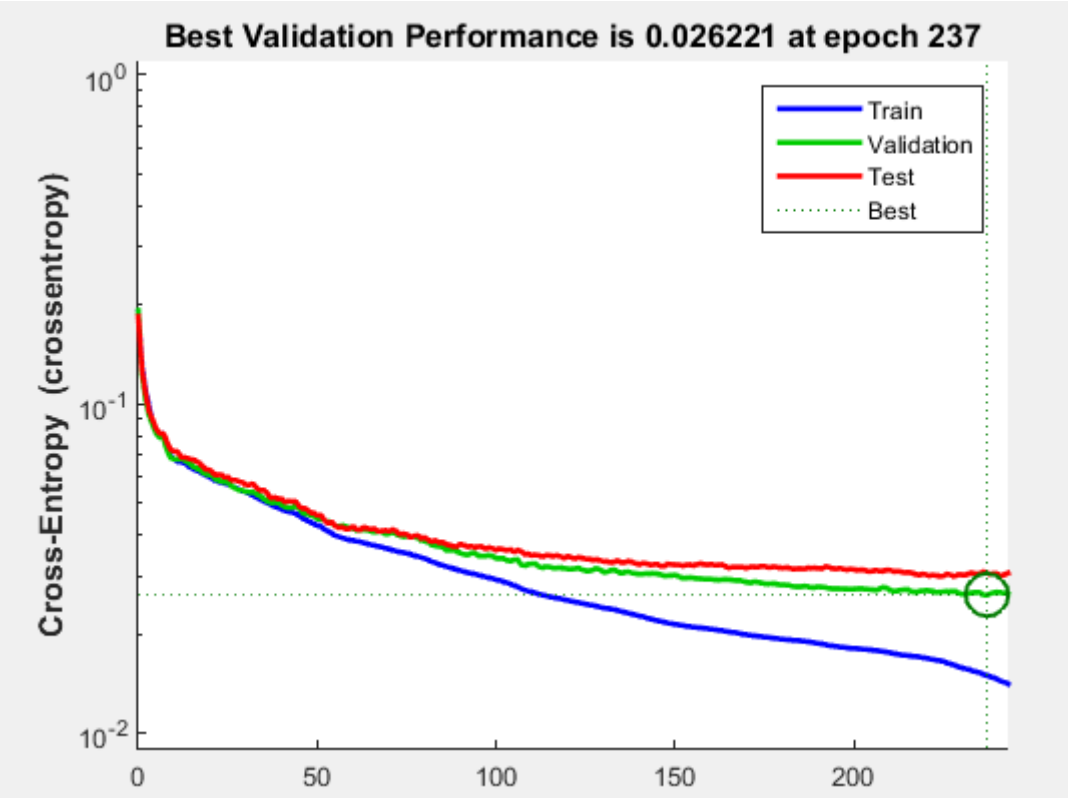
Google otsingumootori trendide ennustused on 1% täpsemad võrrelde twitteri säutsude ennustustega, kuid sarnaselt Twitteri andmetega, keskmine valesti klassifitseerituse protsent on 64%.

Teine uurimisküsimus oli seoses teise neurovõrgu kihi suurusega. Kas hüpotees, rohkem neuroneid on parem, vastab tõele ? Kui klasside arv oli y1 siis keskmine eksimuse protsent vähenes a5 ja a25ga 36%lt 31le protsendile a150 puhul. Kui klasside arv oli 19 siis valesti klassifitseeritud näidete arv langes iga kord kui teise kihi neuronite kogus läks suuremaks. Kõige suuremate klasside arvuga y3 korral oli madalaim vigade arv siis kui neuronite arv a =

25. Seega kolmest katsest kahel tõi neuronite arvu tõus kaasa valesti klassifitseeritud näidete arvu languse. Seega hüpoteesi, et rohkem neuroneid on parem, ümber me ei lükka, leitud tulemused pigem toetavad hüpoteesi.

Kolmandaks eesmärgiks oli, uurida kuidas klasside arvu muutmine mõjutab tulemusi. Selgelt on näha, et kõige väiksem arv klasse on kõige väiksema vigade arvuga. Klasside arvu tõstmine 9lt 19le vähendab ennustuste täpsust. Valesti klassifitseeritud näidete arv tõuseb keskmiselt 34 protsendilt 51 protsendile. Kuid edasine klasside arv tõus 19lt klassilt 33le klassile tõstab ennustuste täpsust. Valed ennustuste arv väheneb 51 protsendilt 48le protsendile. Loogiline oleks eeldada, et klasside tõus tooks alati kaasa rohkem valesti klassifitseeritud näiteid, kuna rohkem klasse nõuavad algoritmilt täpsemat ennustust, kuid antud katse käigus näeme hoopis langust vigade arvus. Kõige suurem vahe klassi y2 ja klassi y3 vahel on kasutades X4 näidete kompleki. Kasutades a25 neuronit, me näeme et y2 klasside korral on ennustute viga 44% ja y3 puhul sama ennustus veaga 18%. Antud fenomenile täpsustavaid selgitusi käesoleva bakalaureusetöö raames ei leitud.

Antud uurimustöö lisaeesmärk oli uurida, kas neurovõrguga treenitud klassifitseerija suudab kõiki neid erinevaid näitajad arvesse võttes, ennustuste täpsuses ületada neid uurimusarendusi, milles on arvestatud ainult sotsiaalmeediast pärit infoga. Nagu eelnevalt mainutud sai, saavutasid (bollen, viide) oma uuringus 86,7 % lise ennustuste täpsuse. Käesoleva bakalaureusetöö märkimisväärseimaks saavutuseks loen fakti, et loodud algoritm suutis ennustada 82 protsendilise täpsusega homset aktsia hinda jäädes +/- 4,09 dollarilistesse piiridesse. Antud meetodit demonstreerime kasutades selleks Tesla Motors Company Inc. (TSLA) aktsiat.



Joonis 10. Plotperform(TR) kuvab treenimise (sinine), rist-validatsiooni (roheline), and testi tulmuste mõõtmise (punane) vead. Mida madalamale langeb joon seda vähem on (mean squared error) valesti klassifitseeritud näiteid. X(4) , y(3) kasutades 25 neuronit.

Kokkuvõte

Käesolev bakalaureusetöö esitab viisi kuidas algoritm õpib edukalt ennustama finantsturgude liikumisi, kasutades sisendina tekste sotsiaalmeediast, trende google otsingumootorist ja mineviku hinnainfot. Antud meetodit demonstreerime kasutades selleks Tesla Motors Company Inc. (TSLA) aktsiat.

Bakalaureusetöö eesmärgiks oli vastata kolmele küsimusele, mis seonduvad turusituatsiooni klassifitseerimise probleemiga.

Esiteks, uuriti milline informatsioon aitab kaasa parema ennustuse tegemisele, ning uurimusarenduse tulemustest võib järeldada, et kõige täpsemaid ennustusi tehakse lähtuvalt andmetest milles on kombineeritud hinnainfo, Tesla CEO Elon Muski säutsud ja Google Tends otsingumootori trendid. Kuid oluline on mainida, et keskmiselt kõige madalamate valede ennustuste arvuga oli hinnainfo andmegrupp.

Teiseks eesmärgiks oli uurida kuidas mõjutab neurovõrgu teise kihi, ehk *hidden layer*i suurus resultaate. Kolmest katsest kahel tõi neuronite arvu tõus kaasa valesti klassifitseeritud näidete arvu languse. Seega hüpotees, mis ütleb, et rohkem neuroneid on parem, leidis pigem kinnitust.

Kolmandaks eesmärgiks oli uurida kuidas klasside arvu muutmine mõjutab tulemusi. Katsetuse tegemiseks kasutati kolme erisuurusega sihtmärk gruppi ehk target classi. Selgus, et suurema täpsusega olid need katsed kus oli kõige vähem klasse. Kuid huvitaval kombel kõige suurema klasside arvuga sihtmärkide vigu oli vähem kui keskmise arvuga target classidel. Antud fenomenile täpsustavaid selgitusi käesoleva bakalaureusetöö raames ei leitud ning selle nähtuse põhjustajaid tasub kindlasti edasi uurida.

Lisa eesmärgina üritasime ületada Bollen, Mao, & Zeng aastal 2011 Twitteri säutsudega Dow Jones aktsia indeksi liikumisi ennustanud uurimusarenduse tulemusi (Bollen et al., 2011). Nende Twitteri säutsude ennustamise täpsus oli 86,7 protsenti. Käesoleva bakalaurese töö märkimisväärseimaks saavutuseks loeksime võimekust ennustada homset Tesla aktsia hinda 82 protsendilise tõenäosusega ning ennustuse täpsusega +/- 4,09 dollarti. Saavutasime ka 96 protsendilise täpsusega ennustuse, kuid see oli küllalti suure, +/- 15,01 dollarilise, sammuga klasside seas. Ennustuse täpsuse kasvu põhjuseks peame Tesla aktsia minevikuhinna lisamist andmete hulka. Kuid selleks, et lisaeesmärki täidetuks lugeda leian, et hinnaliikumise vahemik peaks väiksem olema.

Conclusion

This Bachelors thesis presents a method how an algorithm can successfully learn to predict the movements of a stock price throught the input of social media teksts, trends from the Google search engine and past stock price movements. We demonstrate this methon using the publicly traded Tesla Motors Company inc. (TSLA) stock.

This Bachelors thesis has three main goals that relate to the problem of market situation classification.

First, we research what information, when used as an input, gives most accurate prections. We find that when combined, past price data, Tweets from Tesla's CEO Elon Musk and trends from the Google search engine, we get the most accurate preditctions. It is important to note that on average, the lowest percentage of wrong predictions were made based on historical price data.

The second goal was to find if the amout of neurons on the second layer, i.e. the *hidden layer*, made a differnece to the resuls. In two tests out of three, more neurons on the second layer improved the performance of the algorithm. So the hypothesis that more neurons is better, still stands.

The third goal was to find how changing the amount of target classes influences the predctive performace. Out of three goups of target classes we found the the most accurate was tha target class with the least amount of classes within. Interstingly the target class with the most amount of classes had less misclassified examples than the second biggest tager class. To better understand this phenomena more research need to be done.

As an extra goal we set out to better the performance the Bollen, Mao, & Zeng 2011 Twitter and Dow Jones research paper (Bollen et al., 2011) where they achieved perdiction accuracy of 86,7 percent. This thesis demonstrates a 82 percent probability accuracy within a +/- 4,09 dollar range and a 96 percent probability accuracy within a +/- 15,01 dollar range. We believe that the growth of accuracy came form adding historrical price movements to the dataset. We believe that to consider the extra goal accomplished would be unjust, because of the relatively big price diffenece that it encompasses.

Kasutatud kirjandus

Alanyali, M., Moat, H. S., & Preis, T. (2013). Quantifying the relationship between financial news and the stock market. *Scientific Reports, 3*, 3578. doi:10.1038/srep03578

Anderson, J. A. (MIT press). (1996). An Introduction to Neural Networks | The MIT Press. Võetud 4, 2015, aadressilt http://mitpress.mit.edu/books/introduction-neural-networks

Assis, C. (2015). Elon Musk tweet about "new product line" boosts Tesla shares - MarketWatch. Võetud 1, 2015, aadressilt http://www.marketwatch.com/story/elon-musk-tweet-about-new-product-line-boosts-tesla-shares-2015-03-30

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, 2*(1), 1–8. doi:10.1016/j.jocs.2010.12.007

Boudoukh, J., Feldman, R., Kogan, S., Richardson, M., & Roll, S. (1988).

Da Costa, T. R. C. C., Nazário, R. T., Bergo, G. S. Z., Sobreiro, V. A., & Kimura, H. (2015). Trading System based on the use of technical analysis: A computational experiment. *Journal of Behavioral and Experimental Finance, 6*, 42–55. doi:10.1016/j.jbef.2015.03.003

Duggan, W. (2015). SolarCity And Tesla Have Closer Correlation Than Some Oil Stocks Do With Oil - Tesla Motors, Inc. (NASDAQ:TSLA), (SCTY) | Benzinga. Võetud aadressilt http://www.benzinga.com/general/education/15/04/5373935/solarcity-and-tesla-have-closer-correlation-than-some-oil-stocks-do-

Fama, F. E. (1965). The Behavior of Stock-Market Prices. Võetud 1, 2015, aadressilt http://stevereads.com/papers_to_read/the_behavior_of_stock_market_prices.pdf

Hilbert, M., & López, P. (2011). The world's technological capacity to store, communicate, and compute information. *Science (New York, N.Y.), 332*(6025), 60–65. doi:10.1126/science.1200970

Janssen, C., Langager, C., & Murphy, C. (n.d.). Technical Analysis: Introduction | Investopedia. Võetud April 30, 2015, aadressilt http://www.investopedia.com/university/technical/

Lass, H., & Kippar, J. (2014). Automatiseeritud kauplemissüsteemi optimeerimine FOREX-turu näitel.

LeCun, Y., & Bengio, Y. (1995). Pattern recognition and neural networks. *Handbook of Brain Theory and Neural Networks*, 711. Võetud aadressilt http://books.google.com/books?hl=en&lr=&id=m12UR8QmLqoC&oi=fn

d&pg=PR9&dq=Pattern+recognition+and+neural+networks&ots=aLPmgFUC0e&sig=z8Z37p0es_Q9W8I9N3ZSr0se1g4

Li, Q., Wang, T., Li, P., Liu, L., Gong, Q., & Chen, Y. (2014). The effect of news and public mood on stock movements. *Information Sciences*, *278*, 826–840. doi:10.1016/j.ins.2014.03.096

Li, X., Xie, H., Chen, L., Wang, J., & Deng, X. (2014). News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*, *69*, 14–23. doi:10.1016/j.knosys.2014.04.022

MATLAB. (2013). MATLAB Central. Võetud 3, 2015, aadressilt http://se.mathworks.com/matlabcentral/?s_tid=gn_mlc_logo

Moore, G. E. (1998). Cramming more components onto integrated circuits. *Proceedings of the IEEE*, *86*(1), 82–85. doi:10.1109/JPROC.1998.658762

Ng, A. (2010). Coursera. Võetud 3, 2015, aadressilt https://class.coursera.org/ml-003/lecture

PHP: strpos - Manual. (n.d.). Võetud 3, 2015, aadressilt http://php.net/strpos

Smailović, J., Grčar, M., Lavrač, N., & Žnidaršič, M. (2014). Stream-based active learning for sentiment analysis in the financial domain. *Information Sciences*, *285*(0), -. doi:http://dx.doi.org/10.1016/j.ins.2014.04.034

Zhang, X., Fuehres, H., & Gloor, P. a. (2011). Predicting Stock Market Indicators Through Twitter "I hope it is not as bad as I fear." *Procedia - Social and Behavioral Sciences*, *26*(2007), 55–62. doi:10.1016/j.sbspro.2011.10.562

Veskimägi, M. (2006). EFEKTIIVSE TURU HÜPOTEESI EMPIIRILINE TESTIMINE TALLINNA BÖRSIL. Võetud 1, 2015, aadressilt http://www.nasdaqomxbaltic.com/files/baltic/investor/Mart_Veskimagi.pdf

Lisa 1.

```
<?php
          $a = explode("\n",
file_get_contents('ANDMED_MIDA_SORTIDA_KUUPÄEVA_JÄRGI.txt'));
          $d = explode("\n",
file_get_contents('KUUPÄEVAD.txt'));
          $c= array();
          $f= array();
          for ($x = 0; $x < count($a); $x++) {
               $b = explode("\t", $a[$x]);
               $puhasb = array_map('trim', $b);
               array_push($c, $puhasb);}
          $puhasd = array_map('trim', $d);
          $KontrollDate = "16.04.2015";
          echo "<table>";
          for( $i=0; $i<count($d) ;){
               for( $j=0; $j<count($a);){
               $Date  = ($puhasd[$i]);
                    if ($Date !== "16.01.2010"){
                         $Data = ($c[$j][0]);
                         $pos = strripos($Data, $Date);
                         if ($pos === false) {
                              if($Date == $KontrollDate){
                                   $KontrollDate=$Date;
                                   $i++; } else {
                                   echo "<tr>","<td>",$Date ,
"</td>","</tr>";
                                   $KontrollDate=$Date;
                                   $i++;}
                         } else {
                              if($Date == $KontrollDate){
                                        echo
"<td>",$c[$j][0],"</td>"," ","<td>",$c[$j][1],"</td>";
                                        $KontrollDate=$Date;
```

```
                                        $j++;
                                    } else {
                                    echo "\n", "</tr>", "<tr>" ,
"<td>","$Date", "</td>", " ",
                                        "<td>",$c[$j][0],"
","</td>","<td>",$c[$j][1],"</td>", " ";
                                        $KontrollDate=$Date;
                                        $j++;
                                    }
                                }
                            } else {
                                echo "THE END";
                                echo "</table>";
                                exit;
                            }
                        }
                    } break;
                ?>
```

Programmi koostamisel kasutati AlchemyAPI teenuseid: „www.alchemyapi.com "Text Analysis by AlchemyAPI"“