



**LUNDS UNIVERSITET**  
Ekonomihögskolan

Företagsekonomiska institutionen

Kurskod: FEKH89

Kursens titel: Examensarbete i finansiering på kandidatnivå

Termin: HT2019

## Prestation och Tonläge

En finansiell textanalys av börsnoterade bolag på amerikanska börsen åren  
1994-2018

### **Författare:**

Jesper Carlsson

Jonathan Mathiasson

Robert Srifa

### **Handledare:**

Göran Anderson

# Abstrakt

Titel: Prestation och Tonläge - En finansiell textanalys av börsnoterade bolag på amerikanska börsen åren 1994-2018

Seminariedatum: 16/1-2020

Kurs: FEKH89 - Företagsekonomi: Examensarbete i finansiering på kandidatnivå, 15 högskolepoäng.

Författare: Jesper Carlsson, Jonathan Mathiasson och Robert Srifa

Handledare: Göran Anderson

Nyckelord: Textanalys, bag-of-words, företagsprestation, transparens

Syfte: Studiens syfte är att undersöka om den finansiella prestationen hos företag på den amerikanska börsen mellan år 1994 och 2018 överensstämmer med hur texten i års- och kvartalsrapporter utformats.

Metod: Tonlägesanalysen utgår från metoden bag-of-words och klassificerar ord enligt Loughran McDonald Sentiment Word List. Studien antar en kvantitativ forskningsmetod. Års- och kvartalsrapporter hämtas från EDGAR och den finansiella datan hämtas från COMPUSTAT och CRSP. Data ordnas enligt paneldata med fixed effects i regressionsanalysen.

Teoretiskt ramverk: Transparens, Incomplete hypothesis revelation

Resultat: Urvalet består av cirka 250 000 års- och kvartalsrapporter som sedan matchas med tillhörande finansiell data från COMPUSTAT och CRSP. Studien visar ett positivt samband mellan tonläge och avkastning på tillgångar. Ett positivt samband mellan tonläge och försäljningstillväxt. Ett negativt samband med aktieavkastning.

Slutsats: Det föreligger ett samband mellan finansiell prestation och tonläget i årsredovisningen. Men på grund av väldigt svaga samband dras inga absoluta slutsatser om det härrörs till transparens eller tonlägesmanipulation.

# Abstract

Title: Sentiment and Performance - *A financial text analysis on publicly traded companies on the American stock market years 1994-2018*

Seminare date: 1/16/2020

Course: FEKH89 - Degree Project Undergraduate Level, Business Administration 15 University Credit Points

Authors: Jesper Carlsson, Jonathan Mathiasson och Robert Srifa

Advisor: Göran Anderson

Keywords: Text analysis, bag-of-words, company performance, transparency

Purpose: The purpose of the research is to examine the relationship between sentiment and company performance in 10K and 10Q reports on publicly traded companies on the American stock market during the years 1994-2018.

Methodology: Sentiment analysis made with a bag-of-words method, done with the word classification system Loughran-McDonald Sentiment Word List. 10K and 10Q is retrieved from EDGAR, and financial data is retrieved from COMPUSTAT and CRSP. The study is made with a quantitative approach. Data is sorted by panels and fixed effects are applied to the regression analysis.

Theoretical Perspectives: Transparency, Incomplete Hypothesis Revelation

Empirical foundation: The sample consists of about 250.000 10K and 10Q reports, matched with financial data from COMPUSTAT and CRSP. The empirical results show that there is a positive relation between sentiment and return on assets, as well as between sentiment and sales growth, and a negative relation between sentiment and relative stock return.

Conclusion: The null hypothesis of no relation between sentiment and company performance is rejected. Because of small effect sizes with the sentiment, no absolute conclusions were made with regards to transparency or sentiment manipulation.

# Innehållsförteckning

<b>Abstrakt</b>	<b>1</b>
<b>Abstract</b>	<b>2</b>
<b>Innehållsförteckning</b>	<b>3</b>
<b>1. Inledning</b>	<b>7</b>
1.1 Bakgrund	7
1.2 Problemdiskussion	9
1.3 Syfte	11
1.4 Frågeställningar	11
1.5 Avgränsningar	11
<b>2. Teoretisk referensram och hypoteser</b>	<b>12</b>
2.1 Transparens	12
2.1.1 Incomplete Revelation Hypothesis	13
2.2 Finansiell tonlägesanalys - tidigare forskning	13
2.3 Företagsprestation	16
2.3.1 Return on Assets	17
2.3.2 Avkastning	17
2.3.3 Försäljningstillväxt	17
<b>3. Metod</b>	<b>18</b>
3.1 Tonlägesanalys	18
3.1.1 Bag-of-words	18
3.1.2 Kritik av bag-of-words	18
3.1.2.1 Negationer	18
3.1.2.2 Ordviktning	19
3.1.2.3 Namn	19
3.1.2.4 Tidsaspekten	19
3.1.3 The Loughran-McDonald Sentiment Word List - Finansiellt lexikon	19
3.1.4 Utförande	21
3.1.4.1 Hämta ner års- och kvartalsrapporter från EDGAR	21
3.1.4.2 Beräkna orden i rapporterna	21
3.2 Beroende variabel tonläge	23
3.3 Oberoende variabler - företagsprestation	24
3.3.1 Return on Assets (ROA)	24

3.3.2	Försäljningstillväxt (SALES_GROWTH)	24
3.3.3	Avkastning (RETURN)	24
3.3.4	Kontrollvariabel (LOG_MARKET_CAP)	24
3.3.5	Tidsaspektens effekt på tonläget	25
3.4	Urval	25
3.4.1	Databearbetning - Års- och kvartalsrapporter	25
3.4.2	Databearbetning - COMPUSTAT & CRSP	25
3.5	Regressionsanalys	26
3.5.1	Regressionsmodell OLS	26
3.5.2	Databearbetning - Panelstruktur	27
3.5.2.2	Redundant fixed effects test	27
3.5.3	Regressionsdiagnostik	27
3.5.3.1	Normalfördelade residualer	28
3.5.3.2	Homoskedastisitet	28
3.5.3.3	Multikollinearitet	28
3.5.3.4	Autokorrelation	29
3.5.4	Signifikansnivå - "The large sample fallacy"	29
3.5.5	Winsorizing	29
3.6	Metoddiskussion	29
3.6.1	Reliabilitet	29
3.6.2	Validitet	30
<b>4.</b>	<b>Resultat</b>	<b>32</b>
4.1	Utfall Tonlägesanalys	32
4.1.1	De mest förekommande orden	32
4.1.2	Deskriptiv statistik tonläge	33
4.1.3	Tonläge över tid och industri	35
4.2	Deskriptiv Statistik	36
4.3	Utfall regressionsanalys	37
4.4	Koefficienternas konfidensintervall	38
4.5	Hypotesprövning	39
4.6	Regressionsdiagnostik	39
4.6.1	Medelvärde av feltermen	39
4.6.2	Normalfördelning av feltermen	39
4.6.3	Autokorrelation	40
4.6.4	Multikollinearitet	40
4.6.5	Heteroskedastisitet	41
4.6.6	Hausman-test	42

4.6.7 Redundant Fixed Effect Tests	42
4.6.8 Utelämnade Variabler	42
<b>5. Analys</b>	<b>43</b>
5.1 Analys av deskriptiv data	43
5.2 Analys av regression	43
5.3 Transparens	47
<b>6. Slutsats</b>	<b>49</b>
6.1 Förslag till vidare forskning	49
<b>7. Källförteckning</b>	<b>51</b>
<b>8. Bilagor</b>	<b>57</b>
Bilaga 1: Vanligaste förekommande ord	57
Tabell 8.1: 10 mest förekommande positiva ord	57
Tabell 8.2: 10 mest förekommande negativa ord	57
Bilaga 2. Multipel regressionsanalys	58
Figur 8.1: Multipel regression, nutida mått på företagsprestation	58
Figur 8.2: Multipel regression, framtida mått på företagsprestation	59
Bilaga 3. Koefficienterna konfidensintervall	60
Figur 8.3: Konfidensintervaller, regression med samtida mått på prestation	60
Figur 8.4: Konfidensintervaller, regression med framtida mått på prestation	60
Bilaga 4. Lämplighetstester	61
Figur 8.5. Deskriptiv statistik för feltermerna i den samtida regressionen	61
Figur 8.6: Deskriptiv statistik för feltermerna i den framtida regressionen	61
Figur 8.7: Histogram på feltermerna, regression med samtida mått på prestation	62
Figur 8.8: Histogram på feltermerna, regression med framtida mått på prestation	62
Figur 8.9: Resultat VIF test för regression med samtida prestationsmått	63
Figur 8.10: Resultat VIF test för regression med framtida prestationsmått	64
Figur 8.11: Scatter plot med residualer på y-axeln och estimerat tonläge på x-axeln	64
Figur 8.12: Hausmantest, cross-section random effects, regression med samtida mått på prestation	64

Figur 8.13: Hausmantest, period random effects, regression med framtida mått på prestation	65
Figur 8.14: Hausmantest, cross-section random effects, regression med framtida mått på prestation	65
Figur 8.15: Hausmantest, cross-section random effects, regression med samtida mått på prestation	66
Figur 8.16: Redundant fixed effects test, regression med samtida mått på prestation	66
Figur 8.17: Redundant fixed effects test, regression med framtida mått på prestation	66
Bilaga 5: Pythonkod för tonlägesanalys	67

# 1. Inledning

*I detta kapitel ges en beskrivning på bakgrunden till det valda ämnet som därefter övergår till problemdiskussion. Med hänsyn därtill ges även en redogörelse för syftet med studien och de frågeställningar som studien avser att ge svar på.*

## 1.1 Bakgrund

Efter att James W. Marshall år 1848 fann guld i Sacramento Valley reste tusentals förväntansfulla guldgrävare till Kalifornien i sökandet efter en förmögenhet gömd bland flodbäddarna ("California Gold Rush", 2010). Spade och hacka visar sig vara en ineffektiv metod att hitta guldkorn bland stora mängder sten och sand och tekniker inom hydraulisk gruvsdrift tar snabbt dess plats ("California Gold Rush", 2010). 150 år senare i Silicon Valley några mil söder om Sacramento Valley sitter dataingenjörer och utvecklar nya verktyg för att hitta guldkorn, denna gången bland stora mängder data. Lee (2017) menar på att den tekniska utvecklingen av bland annat web 2.0 (webbplatser som bygger på användarskapat innehåll) och sociala medier har accelererat skapandet av ny data. Trenden har döpts till Big Data och KPMG (2019) anger det som en av de hetaste trenderna inom ny teknologi. Liknelsen med guld är inte allt för långsökt, McKinsey uppskattade det potentiella årliga värdet av Big Data för den europeiska offentliga sektorn till 250 miljarder euro 2011, och McAfee & Brynjolfsson (2012) fann i en studie av 330 Nordamerikanska bolag att företag som anammat trenden har högre produktivitet och lönsamhet.

Vi befinner oss i en tid där data produceras snabbare än någonsin, Wu et al. (2013) menar på att 90 % av all världens data år 2013, ska ha producerats inom loppet av två år. Dessutom framhäver DOMO (2018) i en senare undersökning att det produceras 2,5 kvintiljoner bytes data per dag, vilket motsvarar 2,5 miljarder gigabyte varje dag. Denna extrema mängd data har lett in oss i en era av informationsöverflöd, en era av Big Data (Pan, 2019). Big Data karaktäriseras av *massivitet* (enorma volymer data), *hög hastighet* (snabb informationsbearbetning) och *variation* (olika typer av data) (Pan, 2019).

I ett försök att dra nytta av den enorma mängden data har ämnesområdet "data mining" uppkommit. "Data mining" som är ämnat till att identifiera värdefull information från olika former av datamängder (Pan, 2019). Historiskt sett har data mining präglats av metoder inom artificiell Intelligens, mönsterigenkänning, databassystem och statistik (Kirkos & Manolopoulos, 2004). Vidare förklarar Kirkos



& Manolopoulos (2004) att data mining huvudsakligen används för att upptäcka komplex och ofta dold information från omfattande datamängder. Ett för studiens ämne särskilt talande exempel på tillämpning av Big Data är att U.S. Securities and Exchange Commission (SEC) sedan 2010 använder sig utav den enorma mängd data som genereras i de transaktioner som utförs på aktiebörserna för att hitta anomalier som kan tyda på insiderhandel (Reuters, 2016).

Utvecklingen av en ny generation informationsteknologi har bidragit till att främja många av de möjligheter som den översvämmade informationen medverkat till, däribland subgenren inom data mining som kallas för "text mining" och "text analysis", som avser att hitta mönster inom ostrukturerad data som inte är av numerisk form. Text mining har exempelvis använts för att undersöka den generella atmosfären kring Donald Trumps handlingar genom att analysera information som gjorts tillgänglig via Twitter (Cisija, Zunic & Donko 2018). Teknologier inom textanalys används även av företag i syfte att rikta specifik marknadsföring till sina användare, något som i stor utsträckning sker inom branschen för sociala medier (Cisija, Zunic & Donko 2018). Idén om att analysera text för att identifiera mönster har en lång historia där tidigare studier visat på att T.C Mendall redan år 1901 nyttjat textanalys för att undersöka om vissa verk tillskrivna Shakespeare, istället kan ha skrivits av Bacon (Williams, 1975).

Detta för oss in i den subgenre av textanalys som på engelska heter "*sentiment analysis*". En rak tolkning av ordet *sentiment* är *känsla* men på grund av en avsaknad akademisk svensk term för området och på grund av målet med studien har vi valt att i denna studie översätta detta till "*tonlägesanalys*". Feldman (2013) definierar tonlägesanalys som ett verktyg för att hitta en författares känsla eller åsikt till en viss specifik händelse eller entitet. Han menar på att detta är ett snabbt växande forskningsområde på grund av den tilltagande tillgången av datorkraft som funnits de senaste årtiondena. Han menar att tonlägesanalysen idag bland annat används för att kvantifiera människors uppfattning och åsikter om företag, produkter eller politiker och att detta är en nyckel för att kunna skapa strategier därefter. För att enkelt beskriva tonlägesanalysen handlar det om att klassificera text enligt olika känslor, åsikter eller intentioner och att kvantifiera känslan i texten för att möjliggöra analyser av data i textformat (Feldman, 2013). Text- och tonlägesanalys används också i finansiella sammanhang, där det idag finns bolag som till exempel *Stock Twits* som visar hur ofta en aktie omnämns på twitter, och *The Stock Sonar* som mäter tonläget i börsnoterade bolags omnämnande i finansiell media och visar grafiskt hur tonläget för en viss aktie ändras i jämförelse med aktiens pris.

Inom forskningen har finansiell tonlägesanalys bland annat nyttjats för att mäta och identifiera känslan i bland annat årsredovisningar, kvartalsrapporter, pressmeddelanden och nyhetsartiklar, för att i sin tur kunna förklara en variabel eller hitta tonlägets förklarande variabler. (Tetlock et al., 2008; Feldman et al., 2010; Li, 2010b; Garcia, 2013, Huang, Teoh & Zhang, 2011; Tsai & Wang, 2017; Kang, Park & Han, 2018 Loughran & McDonald, Gandhi, Loughran & McDonald, 2019).

## 1.2 Problemdiskussion

På den amerikanska börsen regleras noterade bolag av United States Securities and Exchange Commission (SEC). Det var efter den stora börskraschen i slutet av 1930-talet som den amerikanska kongressen röstade igenom två lagförslag, Securities Act of 1933 och Securities Exchange Act of 1934, vilket blev starten på den organisation som reglerar börsen idag (U.S. Securities and Exchange Commission, 2013a). Anledningen till införandet är att skydda investerare, förbjuda insiderbrott samt motarbeta nya börskrascher. I och med införandet av Securities Exchange Act of 1934 blev bland annat årsredovisningen en lagstadgad del av aktiebolags, vars tillgångar överstiger \$10 miljoner och vars värdepapper har fler än 500 ägare (United States Securities and Exchange Commission, 2013b), skyldighet att gentemot allmänheten och staten periodvis redovisa sin finansiella ställning. Här lagstadgas det även att denna information ska finnas lättillgänglig för allmänheten (Securities Exchange Act of 1934). Något som ska finnas med i årsredovisningen är den del som härrör bolagens framtidstro, även kallad "Future-looking statement", där investerare ska informeras om framtida resultat och kassaflöden, något som berörs i den del som kallas "Management Discussion and Analysis" (MD&A), där ledningen själva får beskriva med ord hur året har gått och vad framtiden har att erbjuda. Detta regleras bland annat i "Securities Exchange Act of 1934" där det framgår att man inte får vilseleda läsaren (Securities Exchange Act of 1934).

När investerare fattar beslut baseras den på information från kanaler såsom nyheter, pressmeddelanden, analysrapporter, års- och kvartalsrapporter och pressamtal (Kang, Park & Han, 2018). Informationen som investerare erhåller från dessa källor, för att fatta välgrundade beslut uttrycker sig inte bara i numerisk form, utan också i text (Kang, Park & Han 2018). Li (2010a) menar på att text i årsredovisningar innehåller information som är hänförlig till numerisk data samt att den tenderar att fånga upp karaktäristiska egenskaper hos olika befattningshavare. Textens särskiljande egenskaper bidrar således med ett sammanhang som kan vara användbart för att förstå finansiell data och testa ekonomiska hypoteser (Li, 2010a).

Rogers & Grant (1997) redogör för hur ofta finansanalytiker citerar olika delar av årsredovisningen i sina "Sell-side reports" (säljanalyser) och finner att årsredovisningens berättande delar (MD&A, VD ord, styrelseordförande ord och safe harbour statements) i högre grad citeras än sin numeriska motsvarighet (resultatrapport, balansrapport, kassaflödesrapport och fotnoter). De framhäver dock att dessa delar används av finansanalytiker i harmoni med varandra eftersom det som står i de numeriska delarna ofta återges i den berättande texten, men att den berättande texten ibland förtäljer mer än siffrorna. Inom textanalys forskning har man även kommit fram till liknande resultat, där Feldman et al. (2010) finner att tonläget inom MD&A har ett samband med aktiekursförändringen dagen efter årsredovisningen publiceras, vilket talar för att investerare beaktar denna del i sin värdering.

Ghandi, Loughran & McDonald (2019) menar på att de modeller som myndigheter använder sig av för att mäta finansiell nöd inom bankväsendet endast beaktar en kombination av finansiella nyckeltal. De hävdar att problemet med en sådan modell är att befattningshavare, vid vetskap om vilka nyckeltal som särskilt övervakas, kan manipulera dessa i syfte att möta förutbestämda regleringsgränser. De finner att man med hjälp av textanalys kan finna ett samband mellan finansiellt nödställda banker och antalet finansiella ord med negativ laddning i årsredovisningen, något som kan bli en modell i framtida corporate governance metoder. De hävdar att det givetvis går att manipulera tonläget i en årsredovisning, men att risken för stämning bör vara skäl nog att avstå. Å andra sidan hävdar Huang, Teoh & Zhang (2011) att chefer ändrar tonläget i finansiella utlåtanden, när försäljningen varit sämre än väntat, för att dämpa en negativ marknadsreaktion på aktiekursen.

Med beaktning av det informationssamhälle som vi idag lever i, samt den överväldigande mängd data som nu existerar och varje dag skapas kan det vara problematiskt att veta hur denna data kan analyseras och tolkas. Den amerikanska myndighet (SEC) som övervakar bolagens mest fundamentala information arbetar aktivt sedan 30-talet med att tillämpa principer för att underlätta för investerare och aktieägare att få insikt i bolagens prestation, för att låta investerare ta så informerade beslut som möjligt. Bara under 2018 publicerades över 28 000 års- och kvartalsrapporter i SEC:s arkiv, något som talar för den era av Big Data som tidigare har diskuterats. På egen hand torde det vara omöjligt att läsa alla dessa rapporter över en överskådlig tid, men med hjälp av de metoder för textanalys som idag finns och den datorkraft som alla idag har tillgång till, kan detta göras betydligt snabbare.

### 1.3 Syfte

Syftet med denna studie är att undersöka om den finansiella prestationen hos företag på den amerikanska börsen mellan år 1994 och 2018 överensstämmer med hur texten i års- och kvartalsrapporter utformats. Studien kommer att ta hänsyn till den tidsaspekt som texten i års- och kvartalsredovisningen kan avse.

### 1.4 Frågeställningar

Finns det ett samband mellan tonläget i års- och kvartalsrapporter och finansiella mått på företagsprestation under samma period?

Finns det ett samband mellan tonläget i års- och kvartalsrapporter och finansiella mått på företagsprestation från en framtida period?

### 1.5 Avgränsningar

Forskningsfrågorna kommer att avgränsas till att undersöka bolag som lämnar in sina års- och kvartalsredovisningar till SEC och är verksamma på den amerikanska börsen. Tidsserien kommer att utgå från åren 1994 till 2018, och valet av år motiveras av att SEC inte arkiverade rapporterna digitalt förrän år 1994.

Studien avser att mäta sambandet mellan företagsprestation och tonläge och behöver således avgränsas till ett antal mått på prestation. De prestationsmått som kommer tas i beaktning för denna studie är *Return on Assets*, *försäljningstillväxt* och *relativ aktieavkastning*.

För metoden som avser tonlägesanalys avgränsas denne till användandet av två specifika lexikon för klassificering av negativa respektive positiva ord. Dessa lexikon kommer att användas för att beräkna tonläget i års- och kvartalsrapporter.

## 2. Teoretisk referensram och hypoteser

*I detta avsnitt ges en genomgång av tidigare forskning gjord inom ämnet för studien. Sedan presenteras teorier för transparens samt en mothypotes till teorin om effektiva marknader. Sist presenteras en hypotes baserad på den presenterade teorin.*

### 2.1 Transparens

SEC (2018) arbetar aktivt med att förbättra transparensen vid utgivning av finansiell information från börsnoterade bolag till allmänheten, för att skapa en högre grad av symmetri mellan företag, utomstående och marknaden. SEC (2019) har även implementerat och övervakar regeln om "plain-writing" vilket ämnar till att främja transparensen och förståelsen i finansiella rapporter. Samtidigt framhäver SEC även en reform (The Private Securities Litigation Reform Act of 1995), som skyddar offentliga företag från ansvar i privata rättstvister som uppstått på grund av framtidsinriktade uttalanden (t.ex. uppskattningar, prognoser) så länge företaget använt sig av ett försiktigt språk vid uttalandet om framtiden.

Trots att transparens är ett väldiskuterat fenomen inom finansiell rapportering, i samband med en ökad mängd informationsgivning från företag, är begreppet vagt definierat. Blanchet (2002) menar på att transparent information ska härröra till data som används internt för att bedriva verksamheten. Med hänsyn därtill måste den publicerade informationen följaktligen spegla den underliggande ekonomiska verklighet som verksamheten befinner sig i, upplysa om diverse riskfaktorer och möjligheter som verksamheten står inför och slutligen vara förståelig och jämförbar under givna redovisningsprinciper (Blanchet, 2002). Samtidigt lyfter Elorrieta (2002) fram att transparens är starkt förknippat med information av hög kvalitet, som i hennes mening innebär att informationen ska vara trovärdig, vilket kräver att den är granskad enligt hög standard. Transparens berör följaktligen ett flertal koncept: förståelighet, jämförbarhet och trovärdighet.

Några av de studier som gjorts inom ämnet har bland annat undersökt transparensen i MD&A och i pressamtal. Brown och Tucker (2010) undersöker i vilken grad företag ändrar utformningen i MD&A delen från år till år och finner att det är osannolikt att MD&A delen fyller sitt syfte om utformningen inte märkbart förändras i samband med att företagets ekonomiska situation förändras. Mayhew och Venkatachalam (2012) undersöker transparensen i pressamtal och menar på att sättet som befattningshavarna uttrycker sig i innehåller användbar information om ett företags finansiella situation.

### 2.1.1 Incomplete Revelation Hypothesis

I en studie av Fama (1970) konstateras att marknaden är effektiv när priser alltid reflekterar all tillgänglig information. I en studie av Bloomfield (2002) presenteras ett alternativ till Fama (1970) hypotes om effektiva marknader, så kallad "Incomplete Revelation Hypothesis" (IRH). IRH menar på att kostnaden för att utvinna användbar statistik från offentliga uppgifter hindrar marknaderna från att helt avslöja betydelsen av informationen (Bloomfield 2002). Med anknytning därtill betonar Bloomfield (2008) att IRH även indikerar på att chefer kan reducera marknadens reaktion på dåliga nyheter genom att göra dåliga nyheter svårare att analysera. Li (2010a) finner exempelvis att bolag tenderar att göra dåliga nyheter svårare att analysera genom att använda sig av långa meningar eller komplexa ord som inte anses vara nödvändiga i sammanhanget. På liknande sätt menar Bloomfield (2008) att chefer använder sig av mer positiva termer när företag visar på dåliga resultat. Huang, Teoh, Zhang (2011) konstaterar att chefer använder sig av mer positiva termer för att dämpa effekten av dåliga försäljningssiffror vilket tyder på att sambandet mellan marknadens reaktion och textens tonläge möjligtvis kan skapa incitament för chefer att använda sig av en mer positiv ton, än vad som enligt litteraturen anses transparent.

### 2.2 Finansiell tonlägesanalys - tidigare forskning

Inom området för textanalys och inte minst inom den finansiella tonlägesanalysen så diskuterar många forskare om man kan finna det som kallas för "information content" i olika finansiella textkällor (Feldman et al., 2010; Li, 2010b; Loughran & McDonald, 2011). Inom textanalysen så definieras detta begrepp som det ursprungliga meddelandet som den kvantifierade texten är tilltänkt att förmedla, som i sin ursprungliga kontext är lätt att förstå, men som vid en kvantifiering kan vara svårare att hitta (Loughran och McDonald, 2016). Loughran och McDonald (2016) tillägger även att man genom kvantifiering av text också vill finna mönster som talar emot den tilltänkta, ursprungliga avsikten med texten. Kort sagt, innebär "information content" att man i någon utsträckning kan utläsa den ursprungliga intentionen med en text. Detta begrepp har i den tidigare forskningen använts när man hittat starka samband med finansiella variabler (Feldman et al., 2008; Li, 2010; Loughran & McDonald, 2011).

Feldman et al. (2010) gör en tonlägesanalys av MD&A för att se om detta har några direkta eller fördröjda effekter på aktiens avkastning. Syftet med studien är främst att se om texten i MD&A tas i beaktning av investerare när de värderar aktien, och

när detta sker. Tonläget kvantifieras genom att ta skillnaden i positiva och negativa ord genom det totala antalet ord, där orden klassificeras utifrån lexikon Harvard GI (Feldman et al. 2010). De avser att mäta skillnaden i tonläge från en tidigare periods årsrapport, denna variabel antar således ett värde som är positivt om tonläget blivit mer positivt eller negativt om tonläget blivit mer negativt jämfört med perioden innan. Urvalet baseras på alla rapporter som skickats in till SEC från perioden 1994 - 2008 och efter matchning med finansiella variabler så är det cirka 150 000 observationer (Feldman et al. 2010). De finner att när tonläget blir positivare (negativare) så har detta en kortsiktig, signifikant positiv (negativ) effekt på aktieavkastningen. De finner även att tonläget blir positivare när man visar på en god framtida försäljning. Feldman et al. (2010) menar på att investerare använder MD&A delen av årsredovisningen, tillsammans rapportens finansiella data, i sitt beslutsfattande. De framhäver att de krav som SEC ställer på att MD&A måste finnas med, är rättfärdigade då detta förtydligar att kvantifieringen av kvalitativ data (text) innehar så kallad "*information content*".

Huang, Teoh & Zhang (2011) undersöker hur väl tonläget i det som kallas för "earnings press releases" (när bolaget offentligt presenterar hur perioden gått, förkortas till EPR) stämmer överens med de finansiella nyckeltal man baserar dessa utlåtandet på. Syftet med studien är att undersöka om det finns skäl till att tro att befattningshavare strategiskt ändrar sitt tonläge, för att vilseleda investerare om framtida prestationer. Huang et al. (2011) menar på att tonläget är en faktor som investerare beaktar i sin värdering och detta skapar incitament för befattningshavare att manipulera. Studien består av 14 000 observationer, åren 1997-2007, av börsnoterade bolags EPR som sedan matchas med tillhörande finansiell data, där bland annat avkastning på tillgångar, aktieavkastning samt börsvärde är några av tonlägets förklarande variabler. Tonläget kvantifieras genom skillnaden mellan positiva och negativa ord genom den totala mängden ord, orden klassificeras med hjälp av Loughran-McDonald Sentiment Word List. De finner att tonläget i EPR är i genomsnitt positivt, till skillnad från Loughran & McDonalds (2011) studie som analyserar tonläget i årsredovisningar där tonläget i genomsnitt är negativt, något Huang et al (2011) förklarar kan bero på att EPR är en större kanal för att skapa hype bland investerare. Man finner att tonläget i EPR har ett positivt samband med periodens avkastning på tillgångar ( $\beta=0.0011$ ) inget samband med periodens aktieavkastning ( $\beta=0.000$ ) och ett negativt samband med bolagets börsvärde ( $\beta=-0,0002$ ). Det framgår av Huang et al (2011) att ett EPR med en högre andel positiva ord har kortsiktiga effekter på aktiepriset. Dagen efter ett positivt EPR offentliggjorts sker en uppgång i aktiepriset men som efter en kort period återvänder till sin tidigare värdering, något som talar för tonlägets manipulerbarhet

för att vilseleda investerare (Huang et al., 2011). I en liknande studie, men med ett mindre urval (cirka 1400 observationer), så finner Henry (2008) att desto desto positivare tonläget är i en EPR desto högre är den framtida försäljningen uttryckt i siffror.

Li (2010b) undersöker vilka determinanter samt vilket syfte som MD&A har i sitt sammanhang, genom att undersöka vilka finansiella variabler som kan förklara tonläget i den del av MD&A som kallas för "forward-looking statements" (FLS). Genom att undersöka vilka finansiella variabler som kan ha en påverkan på tonläget i FLS och om tonlägets positivitet (negativitet) reflekteras av siffrorna, både i samtid (den period som års- eller kvartalsrapporten avser) och framtid. Li (2010b) mäter tonläget genom den metod som kallas för "Naive-Bayesian Machine Learning", som är en metod baserad på maskininlärning, där ett program själv lär sig vad som uppfattas positivt eller negativt baserat på en mängd inmatade meningar, för att sedan skapa ett kvantifierat mått på tonläge. Studien innehåller cirka 140 000 observationer av börsnoterade bolag, mellan åren 1994-2007. Li (2010b) tror sig hitta ett positivt samband mellan de variabler som i sin natur bör ha ett positivt samband med hur man skriver positivt, det vill säga att högre avkastning på tillgångar/aktie bör göra att befattningshavare skriver positivt, men han reserverar sig med att bestämma samband innan eftersom det enligt tidigare forskning kan finnas tro att utformningen av MD&A kan manipuleras. Han tror på förhand att bolagets börsvärde har ett negativt samband med tonläget, baserat på de politiska kostnader (beskrivs mer ingående i Kap. 3.3.4) som större bolag ofta är associerade med.

Han finner att den samtida avkastningen på tillgångar har ett positivt samband med tonläget ( $\beta=0.293$ ), ett positivt samband med avkastningen ( $\beta=0,047$ ) och ett negativt samband med börsvärdet ( $\beta=-0.003$ ). Liknande samband finner de också med framtidens siffror, där effekten som avkastning på tillgångar och aktieavkastningen har på tonläget är något svagare ( $\beta=0.212$  och  $\beta=0.038$ ) och börsvärdet något starkare ( $\beta=-0.08$ ). Li (2010b) drar inga direkta slutsatser av detta, då det vid tiden för studiens genomförande inte är sagt vilken metod för tonlägesanalys som fungerar bäst för att mäta tonläge, men att man genom denna kvantifiering ändå kan utvinna någon slags ursprunglig intention ("information content") genom kvantifieringen av denna del av rapporten, då inget av sambanden talade emot det man tidigare trott.

I en studie gjord av Gandhi, Loughran & McDonald (2019) som avser att undersöka om man kan identifiera ekonomiskt nödställda banker, genom att mäta sambandet



mellan negativitet i års- och kvartalsrapporten och prestationsmått (med mått på prestation har Return on Assets (ROA) valts som mått). Tonläget räknas ut genom att räkna andelen negativa ord enligt Loughran-McDonalds Sentiment Word List. Gandhi et al. (2019) finner att inom bankbranschen har man en genomsnittlig andel negativa ord på 1,5% i års- och kvartalsrapporten. Resultatet avslöjade ett positivt samband mellan negativare tonläge och lägre ROA.

Kang, Park, Han (2018) avser att mäta sambandet mellan mått på företagsprestation och tonläget i årsredovisningen. Syftet är att testa grundantagandet om att finansiell redovisning skall vara transparent. Tonläget kvantifieras genom skillnaden mellan positiva och negativa ord genom den totala mängden ord, orden klassificeras med hjälp av Loughran-McDonald Sentiment Word List. Urvalet består av cirka 53 000 årsrapporter från börsnoterade bolag, mellan åren 1994 och 2010. Kang et al. (2018) definierar företagsprestation som avkastning på tillgångar. De grundar sin hypotes i att det föreligger ett positivt samband mellan prestation och tonläge. De finner att tonläget är i genomsnitt negativt (-0,8%). Resultatet indikerar att desto högre prestationen är, desto positivare är tonläget (koefficienten för avkastning på tillgångar är  $\beta=0,2890$ ). De finner även ett negativt samband mellan börsvärde och tonläge ( $\beta=-0,0229$ ).

Med stöd av föregående teori och tidigare forskning formuleras följande hypoteser:

**H1:** *Det finns ett samband mellan tonläget i års- och kvartalsrapporter och samtida finansiella mått på företagsprestation*

**H2:** *Det finns ett samband mellan tonläget i års- och kvartalsrapporter och framtida finansiella mått på företagsprestation*

Med samtida finansiella mått avses de mått som rapporteras i samma rapport som tonläget mäts, och framtida finansiella mått avser de mått som avser närmast följande tremånadersperiod.

## 2.3 Företagsprestation

Företagsprestation är ett komplext område att mäta som behandlas på en mängd olika sätt i litteraturen och kan relateras till en variation av faktorer såsom, finansiella nyckeltal, varumärkesvärde, mänskligt kapital eller IT-kompetens (Bharadwaj, 2000; Crook et al., 2011). Med hänsyn därtill finns således inget enhälligt sätt att mäta företagsprestation på. I en studie av Richard et al (2009) konstateras att forskare som avser att mäta prestation måste ta hänsyn till den

multidimensionalitet som råder vid prestationsmätning. Enligt Richard et al. (2009) omfattar företagsprestation huvudsakligen tre specifika områden: finansiell prestation (till exempel: vinst, avkastning på tillgångarna, avkastning på investeringar), prestationer på produktmarknaden (till exempel: försäljning, marknadsandelar) och aktieägarnas avkastning.

### 2.3.1 Return on Assets

Ett flertal studier använder Return on Assets (ROA) som ett mått på företagsprestation (Delen, Kuzey & Uyar, 2013; Rawan, 2019; Naushad, 2019; Usama et al., 2019). Hagel, Brown & Davison (2010) menar på att inget mått i sig är perfekt för att mäta prestation men att ROA däremot kan ge indikationer på hur företag använder sina tillgångar, även Ghandi et al. (2019) använder ROA för att mäta sambandet mellan finansiell prestation och tonläge. ROA avser att mäta hur lönsamt ett företag är relativt sina tillgångar och beräknas genom att dividera företagets nettoinkomst med företagets totala tillgångar. Inom tonlägesanalys har detta nyckeltal används för att beskriva företagsprestation (Li, 2010b; Kang, Park & Han, 2018; Gandhi, Loughran & Mcdonald, 2019).

### 2.3.2 Avkastning

För att fånga upp företagsprestationens multidimensionalitet så är aktieavkastningen ett viktigt nyckeltal för att göra just detta (Richard et al. 2009). Murphy (1999) menar på att avkastningen är en drivande faktor i befattningshavares beteenden eftersom den kan vara ett avgörande mål för bonus. Han menar på att pressen från externa och interna aktieägare, samt pressen att nå sin bonus, gör att befattningshavare tar beslut med detta nyckeltal i sin beaktning. Inom tonlägesanalysen finner Li (2010) ett positivt samband mellan aktiens avkastning per kvartal och tonläget i MD&A. Feldman et al. (2010) finner att desto positivare man är i MD&A desto högre kommer också att avkastningen vara.

### 2.3.3 Försäljningstillväxt

Tillväxten i försäljning avser att mäta den dimension av prestation som befinner sig på produktmarknaden (Richards et al. 2009). Murphy (1999) menar, precis som han argumenterar för avkastningen, att tillväxten i försäljning skapar incitament för ett visst beteende hos befattningshavare. Inom text analysen menar Henry (2008) på att det finns ett positivt samband mellan hög försäljningstillväxt, och en högre andel positiva ord.

## 3. Metod

*Under detta avsnitt ges en redogörelse för vilken metod som använts för tonlägesanalysen samt kritik till denna. Kort därefter kommer en beskrivning av studiens urval, metoden för beroende variabeln och motivering av valda oberoende variabler. Sedan kommer en beskrivning av regressionsanalys samt vilka lämplighetstester som görs och vilken metod analysen av denna del följer. Sist diskuteras helhetsmetoden av studien i termer av reliabilitet och validitet.*

### 3.1 Tonlägesanalys

#### 3.1.1 Bag-of-words

Den vanligaste modellen som används inom tonlägesanalys är bag-of-words modellen (Cambria et al. 2017). Namnet bag-of-words kommer från dess enkla metodik, att enbart räkna antalet ord som uttrycker en specifik känsla eller värdering, där ordföljden ignoreras helt (Cambria et al. 2017). Hur framgångsrik modellen är beror därför i stor utsträckning på hur väl man lyckats skapa associationer mellan ord och känslor, dessa associationer är sparade i ett lexikon som ofta är manuellt framtaget (Cambria et al. 2017). Metoden beaktar heller inte strukturen i texten, utan mäter endast hur ofta ord av samma klassificering förekommer i texten. Kvantifieringen kan på kort tid bidra till en förståelse för tonläget i stora mängder text. Även om den direkta kontexten (meningen, stycket) förbises med denna metod kan man ta hänsyn till den större kontexten (ämnet för texten), genom ett medvetet val av det lexikon som används.

#### 3.1.2 Kritik av bag-of-words

##### 3.1.2.1 Negationer

På grund av den enkelhet som bag-of-words metoden innefattar så behövs metoden anpassas för problematiken kring negerande ord som kan få en annars negativ ordföljd att uppfattas som positiv (till exempel: "not beneficial"). Detta hade enligt metoden räknat ordet beneficial som ett positivt ord trots att den tydligt beskriver något som icke-gynnsamt. Modellen anpassas för detta genom att räkna alla positiva ordföljder som har negationer (no, not, none, neither, never och nobody) inom fyra ord innan det positiva ordet som en negativ ordklassificering (Fin-Neg). Detta görs i enlighet med ordlistans upphovsmän och deras metod för "simple negation" (Loughran & McDonald 2011). I vissa fall kan det negerande ordet

komma efter ordet som det refererar till (till exempel "improvements were never made"), dessa tillfällen har inte tagits hänsyn till och kommer följaktligen att felkategoriseras av metoden, de antas dock vara relativt sällsynta.

### 3.1.2.2 Ordviktning

Loughran och McDonald (2011) betonar att deras lexikon saknar viktning, vilket innebär att listorna inte tar hänsyn till vilka ord som bör väga tyngre än andra utan endast tar hänsyn till antalet ord i varje ordklassificering. Ord som till exempel *Fraud* och *Loss* där det ena i sammanhanget troligtvis kan tolkas värre än den andre kommer att bidra med samma värde i slutändan. Något som även missas i metoden är när det kommer så kallade "modifiers" som vrider tolkning av ett positivt eller negativt ord till att bli mindre eller mer positivt eller negativt. Ord som till exempel *very*, *possibly* eller *pretty* gör att tolkningen av det påföljande ordet förändras, även om tolkningen inte ändrar kategori så finns det en uppenbar skillnad i till exempel "somewhat effective" och "very effective", något som denna metod inte fångar.

### 3.1.2.3 Namn

Ett annat problem som Loughran och McDonald (2011) lyfter fram är att modellen fångar upp vissa namn som positivt eller negativt oavsett om namnet, i sammanhanget, inte avser att förmedla en direkt känsla. Företag med ett värdeord i namnet som till exempel "Best Buy" eller "Innovative Corp" kommer med metoden att få ett uppblåst mått på tonläge om de nämner namnet på företaget ofta. En metod för att motverka denna problematik är att använda sig av skillnaden i tonläge från föregående period istället för tonläge, om namn som finns med i lexikonet kan antas nämnas lika frekvent över tid så tar dessa ut varandra i beräkningen av skillnad i tonläge.

### 3.1.2.4 Tidsaspekten

Något som på förhand blir en tydlig brist med denna metod när man ska mäta det kvantifierade värdet av text med finansiella prestationer, är att man inte genom tonlägesanalysen kan veta när man pratar positivt eller negativt om något, tiden står således stilla i det värde man ger en text. Li (2010b) menar på att man måste testa fler än bara en tidsperiod om man vill finna något av värde, då man i sin tur kan mäta dessa mellan varandra och se vilken av de som har störst effekt.

## 3.1.3 The Loughran-McDonald Sentiment Word List - Finansiellt lexikon

Ett historiskt sett mycket använt lexikon för ordklassificering inom textanalys, är "Harvard Psycho-Sociological Dictionary", även kallat "H4N list of negative words"

(H4N) (Loughran & McDonald, 2011). En av fördelarna med att använda ett väletablerat lexikon som detta är att det inte kan manipuleras av forskaren genom att välja vilka ord som skulle uppfattas som negativa eller inte. Problemet med denna ordlista har visat sig vara att den inte tar hänsyn till den specifika dialekt som hör till den vetenskapsgren man undersöker (Loughran & McDonald, 2011). Det viktigaste verktyget vid klassificering av ord är det lexikon man utgår ifrån när man tolkar innebörden av ett specifikt ord (Feldman, 2013).

Loughran & McDonald (2011) finner att 73,8 % av alla ord i H4N lexikonet är ord som i en finansiell kontext inte bör anses vara negativa. Ord som till exempel *tax*, *cost*, *capital*, *board*, *liability*, *foreign* och *vice* är ord som finns med på listan och som är frekventa inom finansiella texter men som egentligen bara beskriver bolagets styrelse, vice-chefspositioner eller skatter. (Loughran & McDonald, 2011) Även ord som *mine*, *cancer*, *crude*, *tire* eller *capital* tas upp som negativa ord i H4N men som i en finansiell kontext sannolikt beskriver en bransch snarare än ett negativt tonläge. Således presenterade Loughran & McDonald (2011) ett nytt lexikon, med 6 ordklassificeringar som är anpassat till en finansiell kontext.

Loughran & McDonalds (2011) lexikon består av 6 ordklassificeringar:

1. Fin-Neg: Negativa finansiella termer (cirka 2 500 ord som till exempel: *deficit* och *default*).
2. Fin-Pos: Positiva finansiella termer (cirka 350 ord som till exempel: *achieve* och *profit*).
3. Fin-Unc: Termer som betecknar osäkerhet, med betoning på allmän uppfattning om oförmåga snarare än osäkerhet kopplat till risk. (cirka 300 ord som till exempel: *appear* och *doubt*).
4. Fin-Lit: Termer som reflekterar en benägenhet till laglig dispyt (cirka 900 ord som till exempel *lawsuit* och *lawyer*).
5. MW-Strong: Termer som uttrycker nivåer av starkt förtroende (cirka 20 ord som till exempel *always* och *definitely*).
6. MW-Weak: Termer som uttrycker nivåer av lågt förtroende (cirka 30 ord som till exempel *maybe* och *possibly*).

### 3.1.4 Utförande

#### 3.1.4.1 Hämta ner års- och kvartalsrapporter från EDGAR

Första steget i tonlägesanalysen är att hämta alla relevanta filer från SEC:s arkiveringstjänst EDGAR. Samtliga dokument av serie 10-K och 10-Q som laddats upp till U.S. Securities and Exchange Commissions sök- och fillagringstjänst EDGAR (Electronic Data Gathering, Analysis, and Retrieval system) mellan 1 januari 1994 och 31 december 2018 laddas ner och sparas lokalt. 10-K och 10-Q dokument och varianter därav (10-K405, 10-KSB, 10-Q/A med mera) är en årsredovisning respektive kvartalsrapport som måste lämnas in till SEC periodvis av företag med mer än \$10 miljoner i tillgångar och vars värdepapper ägs av fler än 500 personer (U.S. Securities and Exchange Commission, 2013). Dokumentets form och innehåll regleras i Securities Exchange Act of 1934.

Loughran och McDonald beskriver tillvägagångssättet för att ladda ner filerna från EDGAR och rensa de på överflödigt data. De har också tillgängliggjort dessa redovisningar på University of Notre Dames hemsida (McDonald, 2019) för icke-kommersiellt bruk. Det är denna data som kommer användas som underlag för tonlägesanalysen. Datamängden utgör strax över 1 miljon kvartals- och årsredovisningar rapporterade under perioden 1994-2018, vilket motsvarar ca 142GB icke-komprimerade textfiler. Textfilerna har rensats från överflödigt material som inte är relevant för studien, så som bilder, grafer, tabeller (McDonald, 2019). Eftersom rapporterna ursprungligen var i HTML format har de också rensats från märkspråk, det vill säga de HTML-taggar som märkt upp och omgärdat texten. Denna rensning är avgörande för effektiviteten i senare databearbetning då den text som återstår bara utgör några procent av det ursprungliga materialet (McDonald, 2019).

I linje med en annan studie för tonlägesanalys (Kang, Park & Han 2018) rensas datamaterialet för ovanligt korta rapporter (färre än 3 000 ord) och för de rapporter som ersatts av en ny (dubletter). Hur många rapporter som försvunnit i urvalet totalt sett presenteras nedan längre fram, i tabell 3.1.

#### 3.1.4.2 Beräkna orden i rapporterna

Med hjälp av Loughran och McDonalds lexikon och textanalys metoden "Bag-of-words" räknas alla ord i varje enskild årsredovisning och kvartalsrapport, och klassificerar sedan de ord som enligt ordlistan bör uppfattas negativa eller positiva i kategorierna Fin-Neg (negativa ord) och Fin-Pos (positiva ord). Detta

åstadkoms med hjälp av programmeringsspråket Python 3.6. Koden från det script som bearbetat filerna och räknat förekomsten av alla de ord som förekommer i ordlistorna har bifogats i bilaga 5. I nästa stycke följer en kortfattad beskrivning av denna kod, i detta stycke kommer ordet sentiment användas istället för tonläge, då det är detta ord som används av Loughran och McDonald samt i koden.

Rapporterna, som sedan tidigare har sparats i textformat, är arkiverade i en mappstruktur som delar in filerna per år och per månad i detta format: */data/{yyyy}/{mm}/{filnamn}.txt*. Filerna i dessa mappar har itererats rekursivt med början i december 2018, innehållet i varje textfil har lästs in en åt gången och sparats i en variabel, texten har vidare delats in i tokens (en sammanhängande sträng av alfanumeriska tecken) med hjälp av RegEx (regular expressions). Dessa har sedan itererats och de tokens som består enbart av siffror, färre än 2 tecken eller inte ingår i Loughran-McDonald Master Dictionary har ignorerats. Övriga tokens har räknat upp variabeln `wCount` med 1, vilket är den variabel som räknar det totala antalet ord i filen. Antalet gånger det enskilda ordet har förekommit i rapporten räknas också upp i `wDict` som är en variabel av datatypen Dictionary, denna datatyp kan sägas omfatta en lista med datapar, en nyckel och ett värde, vilka i detta fallet är ordet respektive antalet gånger ordet förekommit. Sedan kontrolleras om ordet ingår i någon av de sentimentordlistor som Loughran och McDonald har byggt upp, om så är fallet räknas värdet tillhörande den nyckel med namnet på sentimentet upp i `sDict`. Undantaget är för ord som förekommer i sentimentordlistan Fin-Pos, där det först kontrolleras om en negation förekommit bland de senaste tre orden, om så är fallet räknas istället `sDict` upp i dataparet med nyckel `'negative'`. Efter varje fil sparas resultaten i en excelfil.

För att ge ett tidsperspektiv på processen har vi mätt tiden det tar att bearbeta filerna. Ett test på 26 100 filer innehållandes 702 667 357 ord genomfördes på 2 809 sekunder, alltså 9,3 filer per sekund, eller 250 000 ord per sekund. Trots den synbart höga bearbetningshastigheten kan man alltså räkna med att det tar i genomsnitt 31 timmar att analysera samtliga 1 028 674 filer. Tiden det tar att analysera en rapport måste ändå ses som en av textanalysens främsta styrkor.

För att illustrera hur metoden går till inkluderas ett utdrag ur MD&A delen från Procter & Gambles kvartalsrapport från 1994 avseende räkningskvartalet som avslutades 30 september (figur 3.1). Exemplet är taget från en av de mest positiva rapporterna, med ett beräknat tonläge på 1,8%. I figuren har ord som modellen identifierat som positiva markerats med en grön bakgrund, och röd bakgrund för negativa ord. Gulmarkerade ord är negationer och i detta exempel har negationen

“NOT” vänt värderingen av ordet “EASY” från att vara ett positivt ord till att räknas som ett negativt ord. Utdraget illustrerar också två typer av tillfällen där modellen är bristfällig, ordet “DELIGHT” har räknats som ett positivt ord, trots att det i exemplet är en del av ett namn på en produkt som företaget skriver om (Sunny Delight). Det andra exemplet på hur modellen kan förbättras för att mer exakt kvantifiera tonläget i texten är att när företaget skriver “pretty successfully” så värderas “successfully” lika högt som om företaget hade skrivit “very successfully” eller bara “successfully”.

We are also experiencing **good success** with this strategy in our Food and Beverage business. Sunny Delight, acquired in 1989, is a **good** example. This brand has grown from a 23% share of the rapidly growing juice drink market to a 37% share and market **leadership** in its category today. Sunny **Delight** was moved from a regional brand to a national brand in the U.S. with growth stimulated by unique new products, like Sunny **Delight** Plus Calcium and California Style Punch. Since the acquisition, Sunny Delight's volume has tripled. It is now the Company's 10th largest brand in the United States, and many elements of Sunny Delight's **success** are being shared and reapplied to our growing fruit drink business in Europe. At the time that we announced our decision to exit the 100% orange juice business, we said we were going to refocus our strategic effort on fruit drinks, where we felt there was more **opportunity** for product **innovation** and **creative** marketing. That strategic redirection has made the future for us in this business look a **good** deal more promising. Our second growth strategy is the expansion of core businesses into new markets. This has **not** been as **easy** to do as it might sound. Competition is often strongly entrenched and always **defends against** new entries quite vigorously. But we have been **able** to expand our business around the world pretty **successfully** by following two simple rules: First, we only expand brands when they are healthy and growing in their existing markets. What's working there will usually work elsewhere. And second, we try to move fast while the brand is hot.

Figur 3.1 Exempel på metod för ordräkning (Procter & Gamble, 1994)

### 3.2 Beroende variabel tonläge

Variabel för tonläge benämns inom studien för “TONE”. Denna kommer att beräknas enligt följande formel:

$$TONE = \frac{W_{pos} - W_{neg}}{W_{total}}$$

Där  $W_{pos}$  är antalet positiva ord i rapporten,  $W_{neg}$  antalet negativa ord och  $W_{total}$  det totala antalet ord. Detta görs i enlighet med Huang et al. (2013) och Kang et al.



(2018). Detta ger oss ett effektivt sätt att kvantifiera den annars kvalitativa egenskapen tonläge och kunna identifiera en text som positiv,  $TONE > 0$ , eller negativ,  $TONE < 0$ . Måttet anges i procent då det är den andel ord i texten som är positiva minus den andel ord i texten som är negativa.

### 3.3 Oberoende variabler - företagsprestation

Som mått på företagsprestation undersöks följande variabler.

#### 3.3.1 Return on Assets (ROA)

ROA avser att mäta hur lönsamt ett företag är relativt sina tillgångar och beräknas genom att dividera företagets nettoinkomst med företagets totala tillgångar. Med stöd av Li (2010) och Kang et al. (2018) förväntas att ROA ska ha ett positivt samband med tonläget.

$$ROA = \frac{Net\ Income}{Total\ Assets}$$

#### 3.3.2 Försäljningstillväxt (SALES\_GROWTH)

Försäljningstillväxt beräknas genom den procentuella skillnaden i försäljning mot samma räkenskapskvartal föregående år.

$$SALE\_GROWTH = \frac{Revenue_{Q0}}{Revenue_{Q-4}} - 1$$

#### 3.3.3 Avkastning (RETURN)

Aktiens avkastning (utdelning inkluderat) subtraheras med S&P500 avkastning, beräknat kvartalsvis.

$$RETURN = \frac{Price_0 + Dividends}{Price_{-1}} - 1 - Return_{SP500}$$

#### 3.3.4 Kontrollvariabel (LOG\_MARKET\_CAP)

Detta är en kontrollvariabel för bolagets storlek. Watts & Zimmerman (1986) menar på att större bolag i högre grad är associerade med större "politiska kostnader". Detta antagande grundas i att desto större ett bolag är desto synligare är bolaget gentemot allmänheten och desto större är sannolikheten att bolaget blir föremål för myndighetsutövande (Watts & Zimmerman, 1986), vilket kan komma att påverka hur års- och kvartalsrapporter har utformats (Li, 2010). Bolagets storlek har

även en betydande roll i skillnader för bolagets operationella verksamhet och miljön där bolaget är verksamt inom (Li, 2010b). Variabeln logaritmeras i linje med liknande studier gjorda av Kang, Park & Han (2018)

$$LOG\_MARKET\_CAP = LOG(Price_0 \times Shares_0)$$

### 3.3.5 Tidsaspektens effekt på tonläget

För att redogöra för de skillnader som kan spela in vid val av period för prestationsmått, avser denna studie att mäta dessa under:

1. Tremånadersperioden fram till och med sista dagen i räkenskapsperioden.
2. Tremånadersperioden efter sista dagen i räkenskapsperioden.

Detta görs för att texten i års- och kvartalsrapporten avser främst dessa två tidsperioder. En stor del av texten kan antas handla om den räkenskapsperiod som rapporten avser, och en viss del av texten kan också antas handla om den närmaste framtiden.

## 3.4 Urval

### 3.4.1 Databearbetning - Års- och kvartalsrapporter

Urvalet av års- och kvartalsrapporter baseras på den tillgängliga mängd som finns uppladdat digitalt i SEC:s databas EDGAR, där startåret är 1994, och slutåret är 2018.

### 3.4.2 Databearbetning - COMPUSTAT & CRSP

Den historiska finansiella datan hämtas från COMPUSTAT för att sedan matcha ihop denna datan med bolagets tillhörande års- och kvartalsredovisning. För att veta vilken års- och kvartalsredovisning som hör ihop med vilket bolag så hämtas den historiska datan (net income, total assets, revenue från COMPUSTAT och price, shares och return från CRSP) tillsammans med CIK kod. CIK, eller *Central Index Key*, koden är den kod som SEC använder för att identifiera inlämnade dokument i deras databas EDGAR. (U.S. Securities and Exchange Commission, 2019)

Tabell 3.1: Urvalets bortfall

Databortfall	Antal observationer
Nedladdade 10-X rapporter från perioden 1994-2018	1 028 674
- Rapporter med färre än 3 000 ord	157 928
- Rapporter som ej kunde matchas med COMPUSTAT data	275 809
- Rapporter som ej kunde matchas med CRSP data	310 998
- Rapporter som senare ersatts av en ny rapport (dubletter)	24 187
<b>Antal rapporter i regressionen</b>	<b>259 752</b>

Studien börjar med ett dataset på 1 028 674 stycken rapporter. I linje med liknande studier (Audi, Loughran & McDonald, 2016; Kang et al. 2018) rensas datamaterialet för de rapporter som är skrivna med färre än 3 000 ord. Vid matchning av rapporterna mot data från COMPUSTAT försvinner 275 809 observationer. Anledningen till att såpass många försvinner handlar om att den data som hämtas hem måste matchas med en CIK kod för att kunna härröras till rätt årsrapport, något som saknas för viss data i COMPUSTAT. I samband med matchningen med data från CRSP så rensas det för de privata bolag som finns i den ursprungliga datamängden, då detta behandlar data som endast finns tillgänglig på börshandlade bolag, vilket följer studiens syfte. En andel av de bolag som försvann i matchning med COMPUSTAT kan också antas vara privata bolag.

## 3.5 Regressionsanalys

### 3.5.1 Regressionsmodell OLS

Regressionsanalys används inom kvantitativ forskning för att bevisa sambandet mellan en oberoende variabel, och en eller flera beroende variabler. Den beroende variabeln antas vara stokastisk, vilket innebär att man gör ett antagande om att denne beror på slumpen (Brooks, 2014). En regressionsanalys innebär inte att man kan bevisa kausalitet genom att läsa av koefficienterna, men att koefficienterna indikerar på att det finns bevis för ett linjärt samband mellan dem. (Brooks, 2014) En modell för regressionsanalys som är vanligt förekommande inom kvantitativ forskning kallas för *Ordinary Least Squares* (OLS). Modellen estimerar variablernas samband genom att summan av de kvadrerade residualerna minimeras, så att observationernas avstånd från regressionslinjen blir mindre. (Brooks, 2014)

För en regressionsmodell gäller det att förhållandet mellan beroende och oberoende variabler kan uttryckas linjärt i ett diagram (Brooks, 2014). Vidare gäller även ett antal antaganden för att statistisk slutledning ska kunna äga rum, vilket görs genom ett antal lämplighetstester.

Modellen ser ut som följande:

$$TONE = c + \beta_1 \times SALES\_GROWTH + \beta_2 \times ROA + \beta_3 \times RETURN + \beta_4 \times LOG\_MARKET\_CAP + \sum_{i=1}^N c_i + \sum_{i=1}^N p_i + \varepsilon$$

### 3.5.2 Databearbetning - Panelstruktur

Eftersom den inhämtade datan för att mäta samband sträcker sig över två dimensioner, tid och bolag, kommer den att ordnas enligt en paneldatastruktur.

### 3.5.3 Fixed effects

När man ordnat sin data efter en paneldatastruktur så finns det ett antal funktioner för att bättre tolka sina resultat av regressionen. Genom att fånga upp den genomsnittliga variansen som rör sig över tid och över entitet så hålls dessa konstanta (Brooks, 2014). Vilket kommer att göras genom att applicera *time fixed effects* och *cross-sectional fixed effects* på den slutgiltiga regressionsmodellen. Dessa effekter gör att eventuella samband som rör sig över tid och bolag tas hänsyn till i regressionen (Brooks, 2014).

#### 3.5.2.1 Hausman-test

För att veta om *fixed effects* eller *random effects* ska kunna appliceras på modellen så kommer ett *hausman-test* att göras, där nollhypotesen säger att *random effects* är en effektivare metod än *Fixed effects* (Brooks, 2014).

#### 3.5.2.2 Redundant fixed effects test

För att dra några slutsatser om fixed effects ger någon effekt i den slutliga regressionen så kommer ett "*Redundant Fixed Effects test*" att utföras. Där nollhypotesen innebär att fixed effects variablerna är överflödiga.

### 3.5.3 Regressionsdiagnostik

För att avgöra om OLS är rätt modell för att mäta sambanden är det viktigt att regressionsmodellen i enlighet med OLS uppfyller de antaganden som berör den underliggande datan. Ett korrekt resultat förutsätter därmed att OLS antaganden

om den insamlade datan är uppfyllda. Vilket även kommer ha effekt på vikten i de slutsatser som kan dras utifrån resultatet.

### 3.5.3.1 Normalfördelade residualer

För att OLS ska anses som bästa modell för hypotesprövning så ska antagandet om feltermernas normalfördelning vara uppfyllt (Brooks, 2014). Vidare testas detta genom att mäta residualernas skevhet och kurtosis, vilket görs genom att göra ett *Jarque-Bera*-test. Testet kommer att ge oss ett *Jarque-Bera* värde (JB) där ett högre värde också innebär sämre normalfördelning, där nollhypotesen om att feltermerna är normalfördelade förkastas på 5% signifikansnivå. Det föreligger dock en viss problematik med att använda detta värde för bedömning av normalfördelning när urvalet är väldigt stort, då JB beror på urvalets storlek. Det blir istället viktigare att undersöka skevhet och kurtosis och att dessa inte avviker för mycket från sina gränsvärden (0 och 3), samt undersöka fördelningens form i en graf. Ekvationen för att räkna ut JB ser ut som följande:

$$JB = \frac{n}{6}(S^2 + \frac{1}{4}(K - 3)^2)$$

### 3.5.3.2 Homoskedastisitet

Är variansen av variablernas feltermer konstant så är antagandet om homoskedasticitet uppfyllt (Brooks, 2014). Genom att utföra ett *White*-test så finner man om variablernas feltermer är konstanta, eller om variablerna är heteroskedastiska. Med avseende på paneldatastrukturen kommer funktionen *White-Diagonal* att användas, om heteroskedasticitet föreligger.

### 3.5.3.3 Multikollinearitet

För att undersöka om de förekommer hög korrelation mellan de oberoende variablerna (multikollinearitet), undersöks dessa i en korrelationsmatris. En korrelation på 1 eller -1 innebär perfekt multikollinearitet och korrelation på 0 innebär en total avsaknad av multikollinearitet (Brooks, 2014). I praktiken finns inget absolut värde för var gränsen går, men en korrelation nära 0 är eftersträvänsvärt. Resultatet av en hög grad multikollinearitet medför svårigheter att hålla isär effekterna på den beroende variabeln. Ett VIF test görs även för att mäta vilken grad av multikollinearitet som förekommer i regressionsmodellen.

#### 3.5.3.4 Autokorrelation

För att testa om det förekommer autokorrelation mellan variablernas feltermer så kommer ett *Durbin-Watson* test att utföras. Testet ger oss ett *Durbin-Watson* värde (DW), där värdet 2 innebär att det inte förekommer någon autokorrelation (Brooks, 2014). Eftersträvansvärt är att hamna så nära 2 som möjligt.

#### 3.5.4 Signifikansnivå - "The large sample fallacy"

Statistisk sludledning bygger på uppfattning om att acceptera eller förkasta nollhypotesen. Signifikansnivån blir den grad av slump som en är villig att riskera för att förkasta en sann nollhypotes (Brooks, 2014). Generellt ligger denna nivå mellan 1 och 5 procent beroende på graden av signifikans som studien kräver (Brooks, 2014). Problemet vid väldigt stora urval (i vårt fall cirka 260 000 observationer) är att nollhypotesen alltid förkastas eftersom *p-värdena* ofta är mindre än 5 procent trots väldigt liten eller nästintill obefintlig korrelation, något som också beskrivs som "*The large sample fallacy*" (Lantz, 2013). Detta medför att en kvalitativ analys av ett stort urval istället behöver anta en annan slags metod än vad som traditionell hypotesprövning kräver. Lantz (2013) och Lin et al (2013) föreslår att man istället bör fokusera analysen i det som på engelska kallas "*effect size*", de oberoende variablernas effekt (koefficienter) på den beroende variabeln och jämföra dessa med vad som sägs i litteraturen. Vidare föreslår Lin et al. (2013) att man även presenterar de oberoende variablernas konfidensintervall för koefficienten för att göra tolkningen av dess effekter mer överskådlig och mindre absolut.

#### 3.5.5 Winsorizing

För att rensa för extrema uteliggare i variabler med stor variation har winsorizing applicerats. Detta görs även för att göra residualerna mer normalfördelade. Winsorizing på 1 procentsnivån har applicerats på dessa variabler: TONE, ROA, RETURN och SALES\_GROWTH.

### 3.6 Metoddiskussion

#### 3.6.1 Reliabilitet

Som studien lyfter fram i avsnitt 3.5 finns andra ordlistor som avser mäta tonläget i ett dokument, exempelvis H4N. Skillnaden är att orden i denna ordlista klassificerats på ett annorlunda sätt vilket innebär att användandet av ett sådant lexikon skulle generera ett avvikande resultat från vad denna studie kommer fram

till. Det är däremot viktigt att skilja på val och graden av subjektivitet. I fråga om den ordlista som behandlats i denna studie, har tillämpningen av ordlistan skett i enlighet med Loughran och McDonald (2011), varför inga subjektiva bedömningar görs för att komma fram till tonläget i årsredovisningarna. Om en subjektiv tolkning av orden tillämpas, kommer sannolikheten för att ett identiskt resultat ska kunna genereras, vara avsevärt mindre. Studien ger även en beskrivning på från vilka källor datan inhämtats och hur databearbetningen behandlats, vilket också främjar möjligheterna till att kunna återskapa samma resultat. Således görs bedömningen att möjligheterna till att efterlikna studiens resultat är goda, varför studien också erhåller en hög reliabilitet.

### 3.6.2 Validitet

I enlighet med syftet avser denna studie att mäta tonläget i årsredovisningen. För att få en indikation om detta, som även diskuterats i tidigare avsnitt 3.2, har frekvensen av antalet förekommande negativa och positiva ord räknats. Det huvudsakliga problemet med användandet av en sådan metod är det som berör negationer, ordviktning och namn (avsnitt 3.1.2). Dessa problem medför att det kan förekomma störningar vid tolkning av tonläget i de årsrapporter där dessa företeelser uppstår. På så sätt väcks frågor kring hur denna metod faktiskt kan fånga den verkliga känslan i ett dokument och om den är tillräcklig för att kunna dra eventuella slutsatser. Om så inte är fallet, kan inga direkta slutsatser dras. Detta kommer följaktligen ha en stark effekt på studiens validitet, varför resultatet tolkas med försiktighet. För att undvika detta i den mån det går, har modellen anpassats för negationer i enlighet med vad som sägs i avsnitt 3.1.2.1. Eftersom en korrekt tolkning av tonläget även bygger på ordklassificeringarna är detta också något som kan ha en effekt på studiens validitet, vilket är öppet för diskussion. Hur dessa klassificeringar speglar den verkliga känslan blir svårt att utreda, något som heller inte faller inom ramen för studiens syfte. Det är däremot desto viktigare att betona det faktum att ordlistans lämplighet i sammanhanget är relativt utforskat. Trots detta görs bedömningen att "Loughran-McDonald Sentiment Word List" är den mest lämpliga för studiens syfte i jämförelse med andra alternativ som lyfts fram i litteraturen idag.

Vad gäller urvalet i undersökningen beror en större andel av bortfallet (275 809 observationer) på att Compustat i sin databas saknar data från rapporter eller att dem saknar bolagens CIK-kod. Effekten av ett sådant bortfall blir svårt att tolka eftersom observationerna försvunnit på grund av att data om det rapporterade bolagets egenskaper inte finns tillgängliga. Risken med detta är att det finns en bias i urvalet. Vilket isåfall har en påverkan på validiteten i studien.

En annan fråga som väcks är i vilken utsträckning de valda prestationsmåten faktiskt förklarar företags prestation. Mot bakgrund av det som sägs i avsnitt 2.2 görs bedömningen att de valda prestationsmåten vilar på tillräcklig empirisk grund för att kunna tolkas som användbart i förening med studiens syfte.



## 4. Resultat

*Nedan presenteras de resultat som genererats genom datainsamlingen och regressionsanalysen. Därefter ges en redogörelse för resultatet av de lämplighetstester som gjorts för att regressionen ska vara riktig.*

### 4.1 Utfall Tönlägesanalys

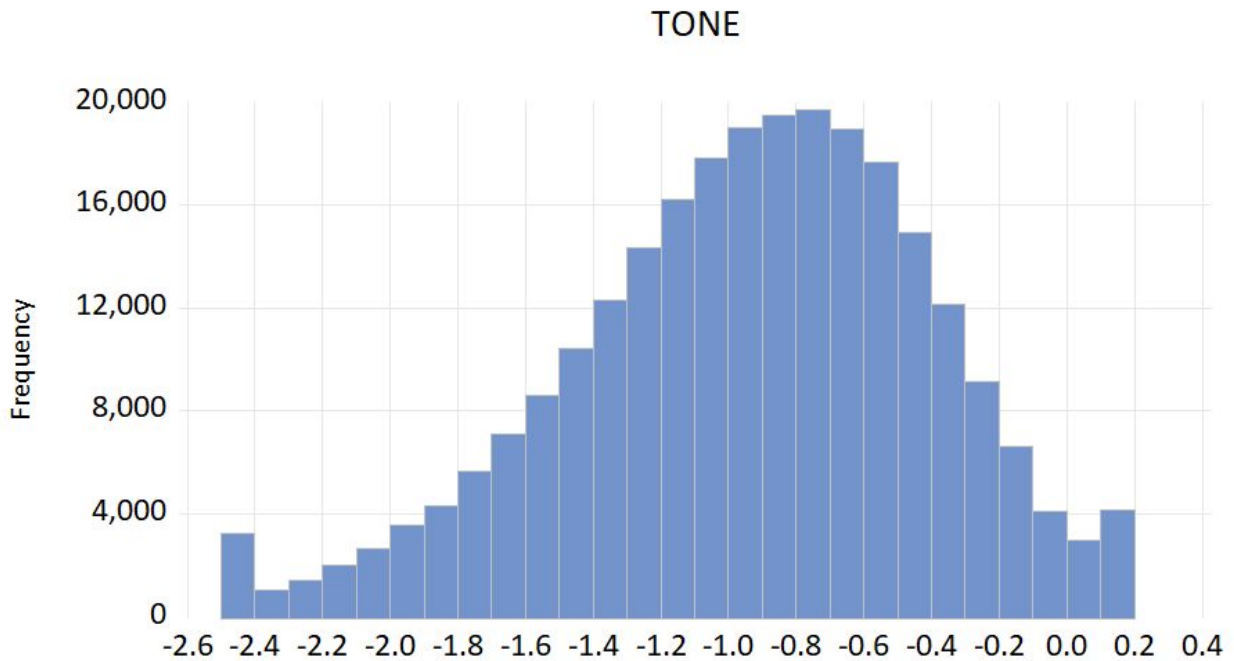
Efter att ha räknat förekomsten av alla ord från Loughran-McDonalds lexikon i samtliga års- och kvartalsrapporter har förekomsten av varje ord summerats kvartalsvis och över hela perioden för att få en övergripande bild av den insamlade datan.

#### 4.1.1 De mest förekommande orden

Figur 4.1 ger en överblick av vilka de mest förekommande positiva orden (i figuren representerad i en grön textfärg) och negativa orden (röd textfärg) är. Textstorleken på orden är indikativt för hur ofta ordet har förekommit. I bilaga 1 bifogas en tabell (8.1) på de 10 mest förekommande positiva orden och hur många gånger de förekommer och en tabell (8.2) med motsvarande negativa ord. Det vanligaste förekommande negativa ordet är *“loss”* och det vanligaste förekommande positiva ordet är *“effective”*.



Figur 4.2. Frekvensen av tonläge vid olika nivåer i års- och kvartalsrapporter 1994-2018. Winzorized på 1 procentsnivån

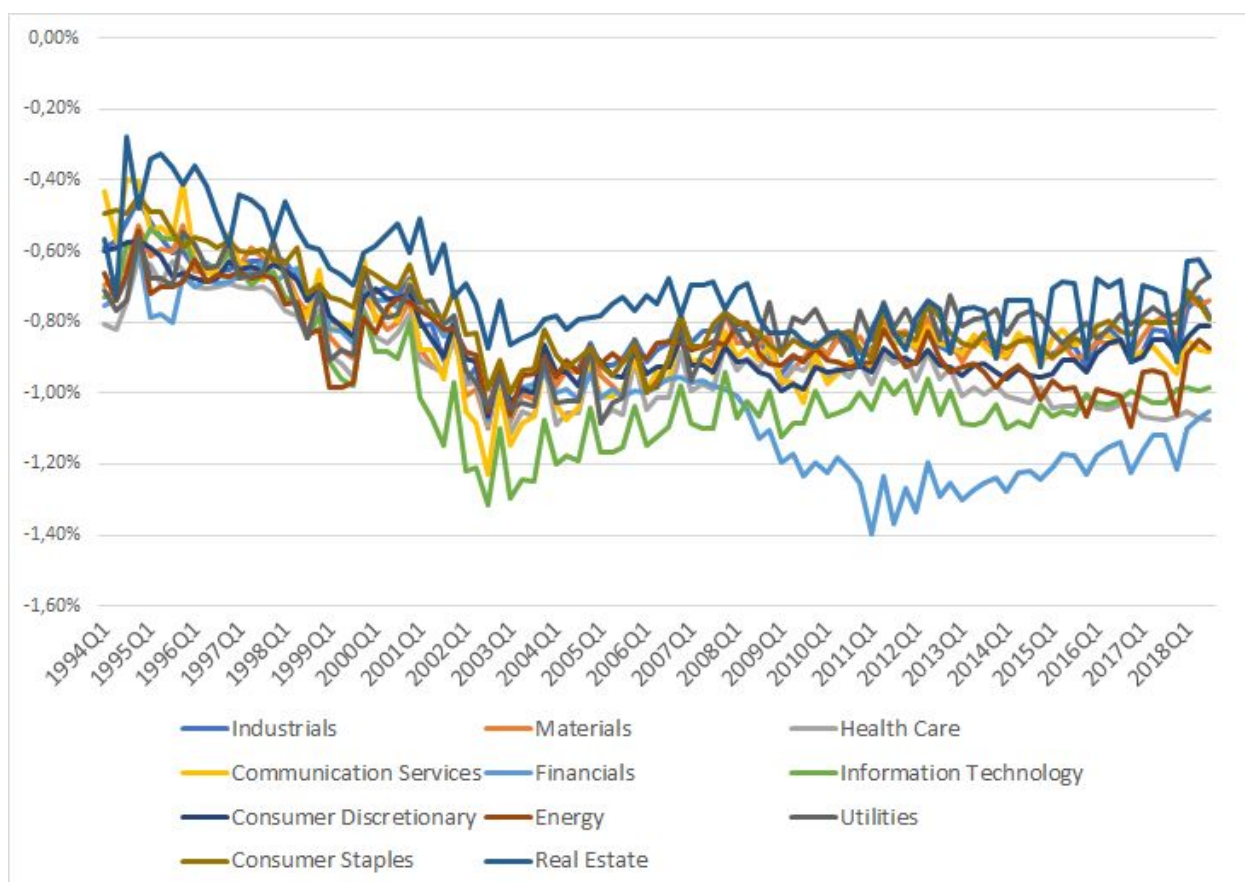


Figur 4.2 visar hur den undersökta variabeln TONE är distribuerad. Endast 2,6% av alla års- och kvartalsrapporter har en högre andel positiva ord i års- och kvartalsredovisningen och därmed ett tonläge på över 0%.

### 4.1.3 Tonläge över tid och industri

Figur 4.3 illustrerar hur tonläget har förändrats över tid inom olika sektorer. Tonläget är beräknat som ett likaviktat genomsnitt av rapporter från företag inom samma sektor, kvartalsindelningen är baserad på datumet för räkenskapsårets slut. Sektorindelningen är baserad på branschklassificeringssystemet GICS som är framtaget av MSCI och S&P, GICS variabeln har matchats ihop med rapporterna med hjälp av data hämtad från COMPUSTAT. Denna graf antyder att tonläget har blivit mer negativt över tid, där det finns avvikande industrier som är generellt sett mer positiva eller mer negativa.

Figur 4.3. Linjediagram på tonläge inom olika sektorer över perioden 1994-2018



## 4.2 Deskriptiv Statistik

Deskriptiv statistik för de oberoende variabler som använts i denna studie presenteras i tabell 4.2. Den visar medelvärdet, medianen, maximivärdet, minimivärdet, standardavvikelse och antalet observationer.

*Tabell 4.2: Deskriptiv statistik för samtliga oberoende variabler. Winzorizing på 1% nivån har applicerats på följande variabler: ROA, RETURN och SALES\_GROWTH*

	Medelvärde	Median	Maximum	Minimum	Standard- avvikelse	Observationer
SALES_GROWTH	19,3011%	7,7600%	439,7544%	-73,4400%	62,0955%	259 752
ROA	-0,4714%	0,4638%	10,7871%	-31,1713%	5,5974%	259 659
RETURN	1,2539%	-0,3600%	95,8322%	-55,8066%	23,5939%	255 528
LOG_MARKET_CAP	13,0180	12,9606	20,6922	5,3931	2,0586	252 965

### 4.3 Utfall regressionsanalys

Utifrån resultatet i båda regressioner konstateras att samtliga oberoende variabler har positiva betakoefficienter, i både samtida- och framtid prestationsmått, förutom RETURN som visar på negativa betakoefficienter i båda fallen.

Signifikansen för respektive oberoende variabel i samtida prestationsmått varierar mellan 97,38% - 100% där return visar på lägst signifikans. I modellen för framtida prestationsmått erhåller alla oberoende variabler trestjärnig signifikans.

Förklaringsgraden för modellerna ligger på cirka 43% (samtida mått) och cirka 44% (framtida mått).

Tabell 4.3: Utfallet av betakoefficienter i regressionerna

	Samtida prestationsmått	Framtida prestationsmått
SALES_GROWTH	0,000190 ***	0,000150 ***
ROA	0,007314 ***	0,004258 ***
RETURN	-0,000087 *	-0,000196 ***
LOG_MARKET_CAP	0,050584 ***	0,058319 ***
INTERCEPT	-1,605712 ***	-1,722054 ***
ADJUSTED $R^2$	0,434914	0,437388
* = signifikant på 5% nivån ** = signifikant på 1% nivån *** = signifikant på 0,1% nivån		

## 4.4 Koefficienternas konfidensintervall

Figur 4.4: Koefficienternas konfidensintervall i regressionen med mått på samtida prestation



Figur 4.5: Koefficienternas konfidensintervall i regressionen med mått på framtida prestation



## 4.5 Hypotesprövning

Samtliga testade variabler visar på statistisk signifikans, det vill säga ett p-värde lägre än 0,05. I den ovan presenterade regressionsanalysen (tabell 4.3) överskådas effekten av varje variabel, där ROA och börsvärde är de två variabler med störst påverkan på tonläget. Alla variabler har en 3-stjärnig signifikans utom den samtida avkastningen (Return) som är av enstjärnig signifikans. Således förkastas nollhypotesen om att inget samband föreligger i både hypotesen som rör de samtida måtten, och i hypotesen som rör de framtida måtten. De slutgiltiga regressionsmodellerna kommer att se ut som följande:

$$TONE_{current} = -1,605712 + 0,000190 \times SALES\_GROWTH + 0,007314 \times ROA + \\ -0,0000874 \times RETURN + 0,050584 \times LOG\_MARKET\_CAP + \sum_{i=1}^N c_i + \sum_{i=1}^N p_i + \varepsilon$$

$$TONE_{past} = -1,722054 + 0,000150 \times SALES\_GROWTH + 0,004258 \times ROA + \\ -0,000196 \times RETURN + 0,058319 \times LOG\_MARKET\_CAP + \sum_{i=1}^N c_i + \sum_{i=1}^N p_i + \varepsilon$$

## 4.6 Regressionsdiagnostik

### 4.6.1 Medelvärde av feltermerna

Medelvärdet av feltermerna är väldigt nära noll i båda regressioner,  $1,81e-17$  för regressionen med samtida mått på prestation respektive  $-6,12e-18$  för regressionen med framtida mått på prestation. Mer deskriptiv statistik för feltermerna är bifogat i bilaga 4 figur 8.5 och 8.6.

### 4.6.2 Normalfördelning av feltermerna

Regressionernas feltermerna har testats för normalfördelning med hjälp av Jarque-Bera och histogram grafer som presenteras i bilaga 4 figur 8.7 och figur 8.8. Utifrån resultatet konstateras att residualerna inte är normalfördelade, och att sannolikheten för normalfördelning är noll i båda regressioner. Skevheten som ska vara 0 vid en normalfördelning  $-0,32$  respektive  $-0,33$ , alltså en viss skevhet i den vänstra svansen. Kurtosis (toppigheten) ska vara 3 vid normalfördelning och är  $4,24$  i båda fallen, alltså leptokurtos. Dessa värden antyder att fördelningen inte är långt ifrån en normalfördelning, men med det stora antalet observationer kan det ändå med säkerhet konstateras att residualerna inte är normalfördelade. Jarque-Bera



värdena är väldigt höga, cirka 18 600 respektive 20 600, vilket inte är oväntat då formeln använder antalet observationer ( $n$ ) som multiplikator.

$$JB = \frac{n}{6}(S^2 + \frac{1}{4}(K - 3)^2)$$

#### 4.6.3 Autokorrelation

För att testa för autokorrelation har ett Durbin-Watson (DW) test genomförts. Resultatet visar att DW-värdet på 1,32 är mindre än det lägsta kritiska värdet på 1,99. Värdet måste falla inom intervallet 1,99-2,01 för att nollhypotesen ej ska förkastas vid ett så här stort urval. Med detta snäva intervall är det alltså inte oväntat att nollhypotesen om ingen autokorrelation förkastas. Det stora urvalet gör att sannolikheten för ett signifikant resultat blir väldigt stor och det är därför mer relevant att se hur stor autokorrelationen är. Detta har gjorts på första, andra och fjärde ordningens fördröjning av residualerna. Fjärde ordningens fördröjning har särskilt valts ut då den avser samma period (kvartal) föregående år. Korrelationen med residualen föregående kvartal är 0,32, korrelationen med residualen två kvartal tidigare är 0,27, och korrelationen med residualen från samma kvartal föregående år är 0,40. Utifrån detta dras slutsatsen att autokorrelation finns och att den är måttlig. Autokorrelationen mellan residualerna och residualerna från samma kvartal föregående år är större, vilket antyder att det finns återkommande tidseffekter som inte fångas upp av modellen.

#### 4.6.4 Multikollinearitet

Enligt korrelationsmatrisen visar samtliga oberoende variabler en låg korrelation med varandra, vilket antyder på att det inte råder hög multikollinearitet. Högsta korrelationen uppmäts mellan ROA och börsvärde med en positiv korrelation på 0,27. Att det finns små skillnader mellan regressionerna beror på att regressionen med framtida prestationsmått använder sig av en släpande beroende variabel vilket gör att en andel av observationerna försvinner i modellen.

Tabell 4.4: Korrelationen mellan variablerna i regressionen med samtida prestationsmått

	TONE	SALES_GROWTH	ROA	RETURN	LOG_MARKET_CAP
TONE	1				
SALES_GROWTH	0,035217	1			
ROA	0,114077	-0,038809	1		
RETURN	0,005008	0,034767	0,106194	1	
LOG_MARKET_CAP	-0,014343	0,003410	0,271849	0,096653	1

Tabell 4.5: Korrelationen mellan variablerna i regressionen med framtida prestationsmått

	TONE(-1)	SALES_GROWTH	ROA	RETURN	LOG_MARKET_CAP
TONE(-1)	1				
SALES_GROWTH	0,032314	1			
ROA	0,093525	-0,028493	1		
RETURN	0,001300	0,036572	0,104022	1	
LOG_MARKET_CAP	-0,009270	0,007560	0,274653	0,091286	1

Resultaten från VIF testerna presenteras i sin helhet i bilaga 4 figur 8.9. Samtliga mått i testet ligger nära 1 och visar tydligt att förekommande multikollinearitet i modellen är väldigt låg.

#### 4.6.5 Heteroskedasticitet

White test för heteroskedasticitet har inte kunnat genomföras på paneldatan, en konservativ hållning har därför tagits och heteroskedasticitet antas förekomma i båda regressioner. Grafiska tester har också gjorts vilka indikerar heteroskedasticitet, en graf för detta har bifogats i bilaga 4 figur 8.11. För att motverka detta har robusta standardfel använts i båda regressioner. Denna metod minskar variansen i standardfelen och gör därmed regressionsmodellerna mer effektiva.

#### 4.6.6 Hausman-test

Enligt vad som kan utläsas i figur 8.12 till 8.15 som finns i bilaga 4 så framgår det att nollhypotesen om att random effects är en effektivare metod förkastas, eftersom  $P > 0.05$ , varför fixed effects tillämpas i denna studie.

#### 4.6.7 Redundant Fixed Effect Tests

Resultaten på dessa tester visar otvetydigt att fixed effects inte är överflödigt i någon av regressionsmodellerna, dessa är bifogade i bilaga 4 figur 8.16 och 8.17.

#### 4.6.8 Utelämnade Variabler

En förhållandevis låg förklaringsgrad och ett högt intercept i förhållande till de oberoende variabelernas betakoefficienter antyder att ett större urval av oberoende variabler hade kunnat hjälpa till att förklara tonläget i rapporterna. Period fixed effects har tillämpats för att fånga upp de faktorer som förändras över tid men är gemensamma för alla bolag, så som konjunktursvängningar. Cross section fixed effects har också tillämpats för att försöka fånga upp de faktorer som är specifika för de rapporterade bolagen och som inte förändras över tid, exempel på detta är företagskultur och även det branschspecifika språkbruket.

## 5. Analys

*I detta stycke förs en diskussion av resultatet från kapitel 4, dessa jämförs med resultat från tidigare forskning samt sätts i perspektiv som härrörs till transparens.*

Studien hade ett stort bortfall av observationer vid matchning av rapporterna med data från COMPUSTAT. Hur detta påverkade urvalet är svårtolkat då kunskap om vilken typ av bolag som föll bort saknas. Hade det handlat om ett par hundra bolag hade det varit lättare att veta exakt varför de saknas, men när det blir ett par hundratusen så ändras detta antagande avsevärt. Bortfallet kommer att beaktas i de slutsatser som dras av studien.

### 5.1 Analys av deskriptiv data

I resultatet konstateras att företag i större utsträckning använder sig av negativa termer. Något som även är konsekvent med tidigare forskning (Kang et al. 2018; Loughran & McDonald, 2011). Orsaken till detta är svår att tolka, men en förklaring till detta kan vara att det finns fler negativa än positiva ord i den ordlista som tillämpats för denna studie (2 500 negativa och 350 positiva).

Denna metod för att mäta tonläge kan troligtvis inte säga något om i vilken grad en ensam rapport är negativ eller positiv, utan att jämföra tonlägesmättet med tonlägesmättet på andra rapporter. Om vald metod för tonlägesanalys varit den bästa i sammanhanget är svårt att veta i efterhand, då graden av subjektivitet i analys av text alltid kommer att vara ett bekymmer, eftersom alla människor har en egen uppfattning om vad som är positivt eller negativt.

### 5.2 Analys av regression

Precis som resultatet förtäljer så förkastas nollhypotesen om att inget samband föreligger i både hypotesen som rör de samtida måtten, och i hypotesen som rör de framtida måtten. Det föreligger således ett samband mellan prestationsmått och tonläget i års- och kvartalsrapporter. Något som dock har konstaterats i och med det som studien tidigare berörde angående "The Large Sample Fallacy" är att hypotesprövning vid större datamängder skall göras med all försiktighet då sannolikheten att något blir signifikant ökar i och med att urvalet blir större (Lantz, 2013; Lin et al., 2013). I linje med det som föreslås av Lantz (2013) och Lin et al. (2013) kommer betakoefficienten att analyseras noggrannare i konfidensintervallen.

Vad gäller ROA mätt i samtida prestationer så har denna koefficient ett positivt beta-värde på 0,007314. Vilket kan avläsas som när ROA går upp med 1%-enhet så blir tonläget i årsredovisningen 0,007314%-enheter mer positiv, förutsatt att alla andra variabler hålls konstanta. Det positiva sambandet är något som även Li (2010), Huang et al. (2013) och Kang et al. (2018) har kommit fram till. Värt att nämna i detta fall är att det endast är Kang. et al. som undersöker samma källa av text, där förvisso Li (2010) undersöker "forward-looking statements" i årsredovisningar, så kommer detta, tillsammans med att våra regressionsmodeller se annorlunda ut, ge oss avvikande resultat från denna forskning. Slutledningen som dras av tidigare forskning kommer således att ligga i att undersöka om de gemensamma variabler har samma positiva eller negativa samband med tonläget, vilket de i fallet ROA tycks ha.

Eftersom års- och kvartalsredovisningen skall berätta om både samtida och framtida prestationer så jämförs båda regressionerna för att se vilket mått som har en större effekt. För de framtida måtten på ROA så finner vi ett något svagare samband än i samtida mått, med ett beta-värde på 0,004258. Över konfidensintervallet för ROAs beta-koefficient rör sig denna mellan 0,0065 upp till 0,0081 och för det framtida måttet mellan 0,0034 och 0,0050 (med en 99,9% konfidensgrad). Utifrån detta kan slutsatsen dras att ROA:s effekt på tonläget, alltid är positiv inom detta urval.

Bloomfield (2002) framhäver i sin studie att företag som visar på dåliga resultat även tenderar att använda sig av mer positiva ord. Genom att beakta resultatet för den regression som berör samtid, framgår ett statistiskt signifikant samband mellan return och tonläget. Det som är särskilt intressant att ta hänsyn till mot bakgrund av det som Bloomfield (2002) beskriver, är att beta-koefficienten för variabeln return erhåller ett negativt värde. Detta skulle i sin mening innebära att företag vars avkastning minskar, allt annat lika, kommer få en positiv effekt på tonläget i årsredovisningen. I motsats till vad denna studie kommer fram till finner Huang et al. (2011) att det inte föreligger något samband mellan variabeln return och tonläget i finansiella utlåtanden. Vid en noggrannare analys av konfidensintervallet för variabelns beta, så framgår det att för det samtida måttet för avkastning så rör sig intervallet mellan ett positivt samband och ett negativt samband ( $\beta = -0,00025$  och  $\beta = 0,00005$ ) med en konfidensgrad på 99,9%. För det framtida måttet så finner man endast ett negativt samband inom intervallet. Det är således tvetydigt vilket samband avkastningen egentligen kan ha på tonläget, men att det i sannolikt är ett negativt sådant.

Försäljningstillväxten har ett positivt samband med tonläget, både i det samtida måttet och det framtida måttet. Eftersom års- och kvartalsredovisningen har delar som både diskuterar hur den gångna perioden har gått samt vad man spår om i framtiden är det relevant att diskutera om det finns några större skillnader mellan de olika tidsaspekterna. För det samtida måttet på försäljningstillväxt så rör sig beta mellan  $\beta=0,00014$  och  $\beta=0,00025$ , det har inom urvalet således alltid en positiv effekt på tonläget. För det framtida måttet är sambandet lite svagare och rör sig närmare ett negativt värde än sin samtida motsvarighet, något som talar för det som Feldman et al. (2010) funnit, där tonläget blir positivare när förutspådd försäljningstillväxt ökar.

Kontrollvariabeln, det logaritmerade börsvärdet, har ett positivt samband med tonläget. Den har även den största effekten av alla variabler på tonläget. Detta talar emot det som initialt hävdats om sambandet med denna och tonläge, där Huang et al. (2011), Li (2010) och Kang et al. (2018) alla finner att storleken på bolaget har ett negativt samband på tonläget. Varför de båda regressionerna finner ett positivt samband kan ha att göra med att studien undersöker olika perioder samt i fallet Li (2010) har denna använt en annan metod för tonlägesanalys. Antagandet om att större bolag i högre grad är försiktiga med uttalanden på grund av associationen med "political costs" som Murphy (1986) föreslår, kan i denna studien inte härröras till tonläget.

För att få en bättre uppfattning om hur stora koefficienterna är, och därmed hur stort sambandet är mellan de oberoende variablerna och den beroende variabeln tonläge, så har de standardiserade koefficienterna beräknats. En standardiserad koefficient är måttet på hur många standardavvikelse den beroende variabeln förändras när en oberoende variabel förändras med en standardavvikelse, givet att allt annat hålls konstant. Exemplifierat är standardavvikelsen av ROA 5,5974% och dess betakoefficient är 0,007314, således kan man enligt modellen dra slutsatsen att om ROA går upp en standardavvikelse (5,5974%) så går tonläget upp  $0,007314 \times 5,5974\% = 0,0409\%$ , allt annat lika. Tonläget har en standardavvikelse på 0,5403%, så denna förändring motsvarar  $\frac{0,0409\%}{0,5403\%} = 0,0758$  standardavvikelse. Motsvarande beräkning har gjorts för alla variabler och presenteras i tabell 5.1.

Tabell 5.1: Beräkningar av standardiserade koefficienter.

Variabel	Beräkning	Standardiserad koefficient
<b>Regression med samtida mått på prestation</b>		
SALES_GROWTH	$\frac{0,00019 \times 62,0955\%}{0,5403\%}$	0,0218
ROA	$\frac{0,007314 \times 5,5974\%}{0,5403\%}$	0,0758
RETURN	$\frac{-0,000087 \times 23,5939\%}{0,5403\%}$	-0,0038
LOG_MARKET_CAP	$\frac{0,050584 \times 2,0586}{0,5403\%}$	0,1927
<b>Regression med framtida mått på prestation</b>		
SALES_GROWTH	$\frac{0,000150 \times 62,0955\%}{0,5403\%}$	0,0172
ROA	$\frac{0,004258 \times 5,5974\%}{0,5403\%}$	0,0441
RETURN	$\frac{-0,000196 \times 23,5939\%}{0,5403\%}$	-0,0086
LOG_MARKET_CAP	$\frac{0,058319 \times 2,0586}{0,5403\%}$	0,2222

Något som måste diskuteras i detta sammanhang är att effekten av dessa variabler är väldigt svaga, trots den, i sammanhanget, stora förklaringsgraden. En förändring på en standardavvikelse av ROA (5,6%) resulterar i en förändring på 0,08 standardavvikelser av tonläget i den rapport prestationen presenteras, och för försäljningstillväxten så kan en förändring på en standardavvikelse (62%) förväntas resultera i en förändring på 0,02 standardavvikelser i tonläget, allt annat lika. Kollar man istället på sambandet med framtida prestationsmått så blir dessa mått ännu svagare.

Ett annat sätt att tolka modellernas koefficienter är att jämföra med intercepten, här framgår det att dessa skiljer sig avsevärt. Interceptet för den samtida modellen är  $C=-1,605712$  och för den framtida modellen  $C=-1,722054$ , där den koefficient, som inte är en kontrollvariabel, som har störst potentiell effekt på tonläget är samtida ROA på  $\beta=0,0084$ . Det krävs således en rejäl ökning av samtliga prestationer, givet att börsvärdet hålls konstant för att påverka tonläget i någon märkbar utsträckning. Förklaringen till den svaga effekten hos prestationsvariablerna kan bero på att det föreligger andra effekter som istället rör sig över tid och entitet, eftersom Fixed Year Effects och Cross-sectional Fixed Effects har applicerats. Metoden kan tyvärr inte mäta hur stor effekten av dessa är på

tonläget, men att det finns en stark effekt är troligt. Detta medför att det är väldigt svårt att säga hur mycket som faktiskt beror av bolagens enskilda prestationer eller vad som beror av andra effekter. I figur 4.3 så kan det utläsas att tiden verkar spela en viss roll i hur positiv eller negativ man är, då det generellt sätt över tid har blivit ett negativare i tonläge i rapporterna.

### 5.3 Transparens

Mot bakgrund av resultatet har alla samtida prestationsmått, förutom return ett positivt samband med tonläget. En tolkning av detta skulle vara att tonläget följer samma riktning som försäljningstillväxt, ROA och börsvärde, allt annat lika. Med hänsyn tagen till det som diskuterats i avsnittet för validitet bör däremot ett sådant resultat tolkas med försiktighet. Att försäljningstillväxt och ROA har en positiv betakoefficient kan däremot ge visst stöd till tesen att bolagen i studien är transparenta i sina redovisningar. Hur starkt detta stöd är blir svårt att säga och som tidigare lyfts fram är sambandet svagt för försäljningstillväxt och ROA. Betakoefficienten för försäljningstillväxt och ROA kan däremot tolkas i linje med Blanchet (2002) som menar på att transparent information ska härröra till data som används internt för att bedriva verksamheten. Med hänsyn till Mayhew och Venkatachalam (2008) som menar på att sättet chefer uttrycker sig på i pressmeddelanden innehåller användbar information om ett företags finansiella situation, kommer denna studien även fram till liknande resultat då både ROA och försäljningstillväxt har ett positivt samband i både samtida och framtida mått. Detta skulle betyda att tonläget i årsredovisningen möjligtvis kan ge indikationer på hur företaget presterar eller kommer att prestera i framtiden. Inom forskningsområdet för finansiell tonlägesanalys har det talats om det som kallas för "information content", vilket i sammanhanget innebär att tonläget förklarar en viss variabel, eller att en viss variabel kan förklara tonläget (Feldman et al., 2008; Li, 2010; Loughran & McDonald, 2011), vilket talar för att det kvantifierade måttet på tonläge kan förklara en slags verklighet i sammanhanget. En möjlig anledning till de svaga sambanden skulle kunna vara att bolagen inte bara blir mer positiva när det går särskilt bra, utan också blir mer positiva när det går särskilt dåligt för att kompensera för detta. Men sambanden med verkligheten finns fortfarande och en, förvisso liten, grad av "information content" kan utläsas.

Som tidigare konstaterats har variabeln return ett negativt samband med tonläget. En möjlig förklaring till detta kan ha att göra med "Incomplete Revelation Hypothesis" som antyder på att chefer kan reducera marknadens reaktion på dåliga nyheter genom att göra dessa svårare att analysera (Bloomfield, 2008). Det som blir intressant i detta sammanhang är att det kan skapa incitament för chefer att



uttrycka sig mer positivt när avkastningen minskar eftersom att de vill dämpa effekten av en sådan händelse. Något som går i direkt motsats till vad litteraturen menar på är transparent (Blanchet, 2002; Elorrieta, 2002). En annan möjlig anledning till varför tonläget följer detta mönster i förhållande till avkastning, kan ha att göra med att SEC endast framhäver generella riktlinjer för hur chefer kan uttrycka sig i årsredovisningen, "plain-writing". Problemet med detta blir att den möjligen inte fångar alla aspekter av transparens, vilket ger utrymme för tolkning och manipulation av chefer. Det ska dock tas i beaktning att samtida mått, som tidigare nämnts, inte med säkerhet kan sägas korrelera negativt, men att det finns en sannolikhet att sambandet mellan return och tonläge förhåller sig på det sättet.

## 6. Slutsats

*I detta stycke presenteras de slutsatser som kan dras av analysen och vald metod, samt förslag till vidare forskning.*

Syftet med studien var att mäta sambandet mellan tonläget och finansiella variabler. Utgångspunkten blev således att försöka kvantifiera text i års- och kvartalsredovisningar till ett mätbart mått på tonläge. För att understryka den absoluta fördelen med detta så ligger detta i att en dator kan "läsa" och "bedöma" en rapport betydligt snabbare än en människa kan. Graden av egen subjektivitet i sammanhanget reduceras också, i och med att bedömningen görs utifrån ett externt ordklassificeringssystem, i detta fall Loughran-McDonald Sentiment Word List. Det kvarstår fortfarande problem gällande metoden Bag-of-Words enkelhet, då ord kan ha en väldigt stor skillnad i hur de tolkas, inte minst när de tas ut ur sitt sammanhang.

Något som visade sig vara ett större problem än vad som på förhand hade trots var den stora datamängden och det bortfall som medfördes när rapporterna skulle matchas med finansiell data, vilket kommer att påverka validiteten i den slutledning som gjorts. Det stora materialet resulterade i att den statistiska slutledningen behövde anta en annorlunda form och de lämplighetstester som gjorts för att veta om OLS modellen är korrekt är inte anpassade till denna mängd data.

I analysen konstaterades att det finns ett statistiskt signifikant samband mellan företagets prestation och tonläget i rapporterna de producerar. Även om sambandet är statistiskt signifikant vill vi inte dra några långtgående slutsatser med anledning av det svaga sambandet som regressionsmodellen resulterade i, och med stöd av det som sägs om signifikansprövning med större datamängder. Om sambanden pekar mot en underliggande verklighet blir således svårt att avgöra. Den transparens som framhävs som ytterst viktig vid framställandet av års- och kvartalsrapporter blev alltså svårare att dra några direkta slutsatser om, eftersom sambanden pekade åt olika håll.

### 6.1 Förslag till vidare forskning

Efter att ha undersökt förhållandet mellan tonläget i års- och kvartalsrapporterna och företagets prestation väcks några frågor som önskas lyfta som ett förslag till vidare forskning.

Som studien lyfter fram, finns ett samband mellan tonläget i den periodvisa rapporteringen och företagets prestation. Det som däremot skulle vara intressant att undersöka är hur effekten av tonläget kan skilja sig mellan olika branscher och vilka yttre effekter som kan komma att påverka tonläget, som till exempel konjunktur eller kultur, förslagsvis genom att göra en eventstudie.

## 7. Källförteckning

Audi, R., Loughran, T. and McDonald, B. (2016) 'Trust, but Verify: MD&A Language and the Role of Trust in Corporate Culture', *Journal of Business Ethics*, 139(3), pp. 551–561. doi: 10.1007/s10551-015-2659-4.

Bharadwaj, A. S. (2000) 'A Resource-Based Perspective on Information Technology Capability and Firm Performance: An Empirical Investigation', *MIS Quarterly*, 24(1), p. 169.

Blanchet, J. (2002) Global Standards Offer Opportunity, *Financial Executive*, 18(2), pp. 28–30.

Bloomfield, R. (2008) 'Discussion of "Annual report readability, current earnings, and earnings persistence"', *Journal of Accounting and Economics*, 45(2), pp. 248–252.

Bloomfield, R. J. (2002) 'The "Incomplete Revelation Hypothesis" and Financial Reporting', *Accounting Horizons*, 16(3), pp. 233–243.

Brooks, Chris. (2014) *Introductory Econometrics for Finance*, Cambridge University Press, 2011.

Brown, S. V. & Tucker, J. W. (2011) Large-Sample Evidence on Firms' Year-over-Year MD&A Modifications', *Journal of Accounting Research*, 49(2), pp. 309–346.

Cambria, E., Das, D., Bandyopadhyay, S., Feraco, A. (2017). *A Practical Guide to Sentiment Analysis*, Cham: Springer

Cisija, M., Zunic, E. and Donko, D. (2018) 'Collection and Sentiment Analysis of Twitter Data on the Political Atmosphere', 2018 14th Symposium on Neural Networks and Applications (NEUREL), Neural Networks and Applications (NEUREL), 2018 14th Symposium on,

Crook, T. Russel., Todd, Y. Samuel., Combs, G. James., Woehr, J. David & Ketchen, J. David., (2011) 'Does Human Capital Matter? A Meta-Analysis of the Relationship between Human Capital and Firm Performance', *Journal of Applied Psychology*, 96(3), pp. 443–456.

Domo (2018) Data Never Sleeps 6.0 Tillgänglig online:  
[https://www.domo.com/assets/downloads/18\\_domo\\_data-never-sleeps-6+verticals.pdf](https://www.domo.com/assets/downloads/18_domo_data-never-sleeps-6+verticals.pdf)  
(Hämtad 8 Januari 2020)

Delen, D., Kuzey, C. and Uyar, A. (2013) 'Measuring firm performance using financial ratios: A decision tree approach', *Expert Systems With Applications*, 40(10), pp. 3970–3983.

Elorietta, A. M. (2002) DISCLOSURE AND TRANSPARENCY - Accounting and Auditing, Organisation for Economic Co-operation and Development, Tillgänglig:  
<http://www.oecd.org/corporate/ca/corporategovernanceprinciples/2085716.pdf>  
(Hämtad: 8 januari 2020)

FAMA, Eugene. F. (1970) 'Efficient Capital Markets: A Review of Theory and Empirical Work', *Journal of Finance* (Wiley-Blackwell), 25(2), pp. 383–417.

Feldman, R. (2013) Techniques and Applications for Sentiment Analysis, *Communications of the ACM*, 56(4), pp. 82–89.

Feldman, R., Govindaraj, S., Livnat, J. & Segal, B. (2010), Management's tone change, post earnings announcement drift and accruals, *Review of Accounting Studies*, 15(4), pp. 915–953.

Gandhi, P., Loughran, T. & McDonald, B. (2019) 'Using annual report sentiment as a proxy for financial distress in uS Banks', *Journal of Behavioral Finance*.

García, Diego (2013), Sentiment during Recessions, *The Journal of Finance*, 68(3), p. 1267.

Hagel, J., Brown, S. & Davison, L. (2010) The Best Way to Measure Company Performance, *Harvard Business Review* Tillgänglig:  
<https://hbr.org/2010/03/the-best-way-to-measure-compan.html>

Henry, E. (2008), Are Investors Influenced by How Earnings Press Releases Are Written?, *Journal of Business Communication*, 45(4), pp. 363–407.

History (2010). American Gold Rush: Tillgänglig online:  
<https://www.history.com/topics/westward-expansion/gold-rush-of-1849> (Hämtad 8 Januari 2020)

Huang, Xuan., Teoh, Siew Hong. & Zhang, Yinglei. (2011), Tone Management, *Accounting Review*, 89(3), pp. 1083–1113.

KPMG (2019). Technology Trends Index USA: Tillgänglig online: <http://technologytrendsindex.kpmg.com/> (Hämtad 8 Januari 2020)

Kang, T., Park, D. & Han, I. (2018) 'Beyond the numbers: The effect of 10-K tone on firms' performance predictions using text analytics', *Telematics and Informatics*, 35(2), pp. 370–381.

Kirkos, E. & Manoupolus, Y. (2004) Data Mining In Finance And Accounting: A Review Of Current Research Trends. Tillgänglig online: <http://www.academia.edu/download/30842460/10.1.1.76.7734.pdf> (Hämtad 10 Januari 2020)

Lantz, Björn. (2013) 'The large sample size fallacy', *Scandinavian Journal of Caring Sciences*, 27(2), pp. 487–492. Tillgänglig: Hämtad:

Lee, I 2017, 'Big data: Dimensions, evolution, impacts, and challenges', *Business Horizons*, vol. 60, no. 3, pp. 293–303

Lee, I. (2017) 'Big data: Dimensions, evolution, impacts, and challenges', *Business Horizons*, 60(3), LUSEM Library website <http://www.lusem.lu.se/library> (Tillgänglig: 10 Januari 2020).

Li, F. (2010a) Textual Analysis of Corporate Disclosures: A Survey of the Literature, *Journal of Accounting Literature*, 29, pp. 143-165.

Li, F. (2010b) The Information Content of Forward-Looking Statements in Corporate Filings-A Naive Bayesian Machine Learning Approach, *Journal of Accounting Research*, 48(5), pp. 1049–1102.

Loughran, T. & McDonald, B. (2016) 'Textual Analysis in Accounting and Finance: A Survey', *Journal of Accounting Research*, 54(4), pp. 1187–1230. Tillgänglig: Hämtad:

Loughran, T. & McDonald, B. (2011) When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks, *The Journal of Finance*, 66(1), p. 35.

Mayew, W. J. & Venkatachalam, M. (2012) 'The Power of Voice: Managerial Affective States and Future Firm Performance', *The Journal of Finance*, 67(1), p. 1

McAfee, A. & Brynjolfsson, E. (2012). *Harvard Business Review Big Data: The Management Revolution* [pdf] Tillgänglig via:  
<http://tarjomefa.com/wp-content/uploads/2017/04/6539-English-TarjomeFa-1.pdf>

McKinsey Global Institute (2011). *Big Data: The next frontier for innovation, competition and productivity* [pdf] Tillgänglig via:  
[https://www.mckinsey.com/~media/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Big%20data%20The%20next%20frontier%20for%20innovation/MGI\\_big\\_data\\_full\\_report.ashx](https://www.mckinsey.com/~media/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Big%20data%20The%20next%20frontier%20for%20innovation/MGI_big_data_full_report.ashx) (Hämtad: 7 januari 2020)

McDonald, B. (2019). *Data*. Tillgänglig Online: <https://sraf.nd.edu/data/> [hämtad 14/1-2020]

McDonald, Bill (2019) *Stage One 10-X Parse Data*, Tillgänglig online:  
<https://sraf.nd.edu/data/stage-one-10-x-parse-data/> [hämtad: 9 januari 2020]

Mohammad Naushad (2019) 'Intellectual capital and financial performance of Sharia-compliant banks in Saudi Arabia', *Banks and Bank Systems*, (4), p. 1.

Murphy, K. J. (1999) Chapter 38 Executive compensation, *Handbook of Labor Economics*, Volume 3 pp 2485-2563

Pan, Lili (2019) 'A Big Data-Based Data Mining Tool for Physical Education and Technical and Tactical Analysis', *International Journal of Emerging Technologies in Learning*, 14(22)

Rawan, Abuzarqa (2019) 'Evaluating Banks Financial Performance Using Financial Ratios: A Case Study of Kuwait Local Commercial Banks', *Oradea Journal of Business and Economics*, (2), p. 56.

Reuters (2016) SEC's advanced data analytics helps detect even the smallest illicit market activity, *Reuters*, 1 November, Tillgänglig Online:  
<https://fortune.com/2016/11/01/sec-big-data-insider-trading/> [hämtad 14/1-2020]

Richard, P. J., Devinney, T, Yip, G., Johnsson, G . (2009) 'Measuring Organizational Performance: Towards Methodological Best Practice', Journal of Management, 35(3), pp. 718–804.

Rogers, K., & Grant, J. (1997), Content analysis of information cited in reports of sell-side financial analysts, Journal of Financial Statement Analysis, 3(1), p. 17.

SECURITIES EXCHANGE ACT OF 1934 (2018) U.S. Securities Exchange Commission Tillgänglig online:

<https://legcounsel.house.gov/Comps/Securities%20Exchange%20Act%20Of%201934.pdf> (Hämtad: 10 januari 2020)

Tetlock, P. C., Saar-Tsechansky, M. & Macskassy, S. (2008), More than Words: Quantifying Language to Measure Firms' Fundamentals, The Journal of Finance, 63(3), p. 1437.

The Private Securities Litigation Reform Act. (1995). U.S. Securities Exchange Commission Tillgänglig online:

<https://www.govinfo.gov/content/pkg/PLAW-104publ67/html/PLAW-104publ67.htm> (Hämtad: 10 januari 2020)

Tsai, Ming-Feng. & Wang, Chuan-Ju. (2017), On the risk prediction and analysis of soft information in finance reports, European Journal of Operational Research, 257(1), pp. 243–250.

U.S. Securities Exchange Commission (2013b) The Laws That Govern the Securities Industry Tillgänglig online:

<https://www.sec.gov/answers/about-lawsshtml.html#secexact1934> (Hämtad 10 Januari 2020)

U.S. Securities Exchange Commission (2019) ,CIK,

Tillgänglig:<https://www.sec.gov/edgar/searchedgar/cik.htm>, (Hämtad: 15 december 2019=

U.S. Securities and Exchange Commission, (2013a) What we do, Tillgänglig:

<https://www.sec.gov/Article/whatwedo.html> (Hämtad: 9 Januari 2020)



U.S. Securities and Exchange Commission, (2018) '2018 SEC annual reports - Form 10-K' Tillgänglig:  
[https://www.ey.com/Publication/vwLUAssets/SECAnnualReports10K\\_05131-181US\\_6December2018-v2/\\$FILE/SECAnnualReports10K\\_05131-181US\\_6December2018-v2.pdf](https://www.ey.com/Publication/vwLUAssets/SECAnnualReports10K_05131-181US_6December2018-v2/$FILE/SECAnnualReports10K_05131-181US_6December2018-v2.pdf) (Hämtad: 9 Januari 2020)

U.S. Securities and Exchange Commission, (2019) 'Plain Writing Initiative' Tillgänglig:  
<https://www.sec.gov/plainwriting.shtml> (Hämtad: 9 Januari 2020)

Usama, A. 2019, 'Impact of Corporate Social Responsibility on Financial Performance of Non-Financial Firms: Evidence from Pakistan Stock Exchange', Pakistan Journal of Social Sciences (PJSS), vol. 39, no. 3, pp. 1083–1089

Willimams, C. B. (1975) 'Mendenhall's studies of word-length distribution in the works of Shakespeare and Bacon', Biometrika, 62(1),

Watts, R. L., & Zimmerman J. L.. Positive Accounting Theory. Englewood Cliffs, NJ: Prentice- Hall, 1986

Xindong, Wu., Xingquan, Zhu., Gong-Qing., Wu. & Wei, Ding. (2014), Data mining with big data, IEEE Transactions on Knowledge and Data Engineering, Knowledge and Data Engineering, IEEE Transactions on, IEEE Trans. Knowl. Data Eng, 26(1), pp. 97–107.

## 8. Bilagor

### Bilaga 1: Vanligaste förekommande ord

Tabell 8.1: 10 mest förekommande positiva ord

Ord	Antal
EFFECTIVE	16 688 196
BENEFIT	9 976 159
ABLE	4 167 393
GAINS	4 061 661
GAIN	3 902 584
GREATER	3 617 990
GOOD	3 473 053
BENEFICIAL	2 746 678
BEST	2 715 739
ADVANCES	2 621 782

Tabell 8.2: 10 mest förekommande negativa ord

Ord	Antal
LOSS	18 820 385
LOSSES	12 695 752
TERMINATION	9 847 810
AGAINST	9 579 217
CLAIMS	8 680 265
DEFAULT	7 108 272
CLOSING	6 751 504
IMPAIRMENT	6 729 303
ADVERSE	6 435 684
ADVERSELY	5 793 481

## Bilaga 2. Multipel regressionsanalys

Figur 8.1: Multipel regression, nutida mått på företagsprestation

Dependent Variable: TONE  
Method: Panel Least Squares  
Date: 01/12/20 Time: 19:11  
Sample: 1993Q2 2018Q4  
Periods included: 103  
Cross-sections included: 8443  
Total panel (unbalanced) observations: 248751  
White diagonal standard errors & covariance (d.f. corrected)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
SALES_GROWTH	0.000190	1.64E-05	11.59888	0.0000
ROA	0.007314	0.000250	29.24044	0.0000
RETURN	-8.74E-05	3.93E-05	-2.223545	0.0262
LOG_MARKET_CAP	0.050584	0.001450	34.89743	0.0000
C	-1.605712	0.018910	-84.91376	0.0000

Effects Specification

Cross-section fixed (dummy variables)			
Period fixed (dummy variables)			
Root MSE	0.399322	R-squared	0.454333
Mean dependent var	-0.946658	Adjusted R-squared	0.434914
S.D. dependent var	0.540580	S.E. of regression	0.406366
Akaike info criterion	1.070639	Sum squared resid	39665.37
Schwarz criterion	1.428895	Log likelihood	-124612.2
Hannan-Quinn criter.	1.175092	F-statistic	23.39690
Durbin-Watson stat	1.324828	Prob(F-statistic)	0.000000

Figur 8.2: Multipel regression, framtida mått på företagsprestation

Dependent Variable: TONE(-1)  
 Method: Panel Least Squares  
 Date: 01/12/20 Time: 19:28  
 Sample (adjusted): 1993Q4 2018Q4  
 Periods included: 101  
 Cross-sections included: 8024  
 Total panel (unbalanced) observations: 229649  
 White diagonal standard errors & covariance (d.f. corrected)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
SALES_GROWTH	0.000150	1.77E-05	8.486844	0.0000
ROA	0.004258	0.000251	16.98723	0.0000
RETURN	-0.000196	4.10E-05	-4.788762	0.0000
LOG_MARKET_CAP	0.058319	0.001544	37.77736	0.0000
C	-1.722054	0.020216	-85.18433	0.0000

Effects Specification

Cross-section fixed (dummy variables)  
 Period fixed (dummy variables)

Root MSE	0.395277	R-squared	0.457299
Mean dependent var	-0.958982	Adjusted R-squared	0.437388
S.D. dependent var	0.536565	S.E. of regression	0.402463
Akaike info criterion	1.052326	Sum squared resid	35881.24
Schwarz criterion	1.418443	Log likelihood	-112704.8
Hannan-Quinn criter.	1.159439	F-statistic	22.96803
Durbin-Watson stat	1.325758	Prob(F-statistic)	0.000000

## Bilaga 3. Koefficienterna konfidensintervall

Figur 8.3: Konfidensintervaller, regression med samtida mått på prestation

Coefficient Confidence Intervals  
Date: 01/12/20 Time: 18:44  
Sample: 1993Q2 2018Q4  
Included observations: 248751

Variable	Coefficient	90% CI		95% CI		99% CI		99.9% CI	
		Low	High	Low	High	Low	High	Low	High
SALES_GROWTH	0.000190	0.000163	0.000217	0.000158	0.000223	0.000148	0.000233	0.000136	0.000245
ROA	0.007314	0.006902	0.007725	0.006824	0.007804	0.006670	0.007958	0.006491	0.008137
RETURN	-8.74E-05	-0.000152	-2.28E-05	-0.000164	-1.04E-05	-0.000189	1.39E-05	-0.000217	4.19E-05
LOG_MARKET_CAP	0.050584	0.048200	0.052968	0.047743	0.053425	0.046850	0.054318	0.045814	0.055354
C	-1.605712	-1.636817	-1.574608	-1.642775	-1.568649	-1.654422	-1.557003	-1.667937	-1.543488

Figur 8.4: Konfidensintervaller, regression med framtida mått på prestation

Coefficient Confidence Intervals  
Date: 01/12/20 Time: 18:45  
Sample: 1993Q2 2018Q4  
Included observations: 229649

Variable	Coefficient	90% CI		95% CI		99% CI		99.9% CI	
		Low	High	Low	High	Low	High	Low	High
SALES_GROWTH	0.000150	0.000121	0.000179	0.000115	0.000185	0.000105	0.000196	9.19E-05	0.000208
ROA	0.004258	0.003846	0.004670	0.003767	0.004749	0.003612	0.004904	0.003433	0.005083
RETURN	-0.000196	-0.000264	-0.000129	-0.000277	-0.000116	-0.000302	-9.07E-05	-0.000331	-6.14E-05
LOG_MARKET_CAP	0.058319	0.055780	0.060859	0.055294	0.061345	0.054343	0.062296	0.053239	0.063399
C	-1.722054	-1.755306	-1.688802	-1.761676	-1.682432	-1.774127	-1.669982	-1.788575	-1.655533

## Bilaga 4. Lämplighetstester

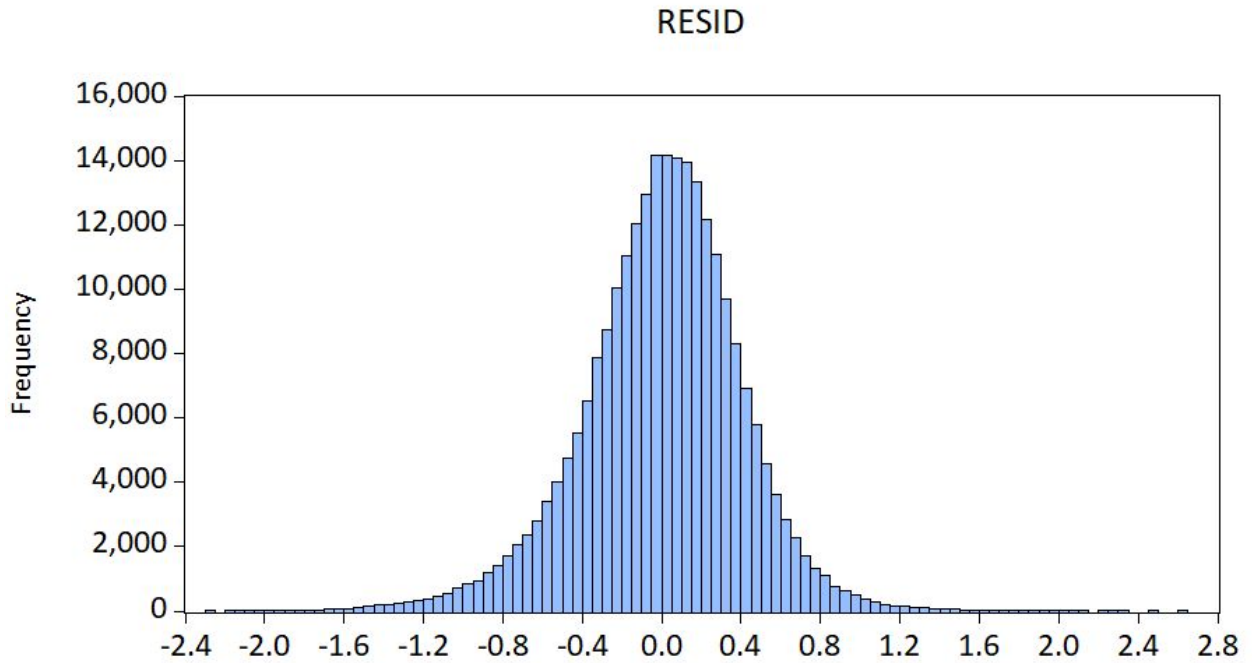
Figur 8.5. Deskriptiv statistik för feltermerna i den samtida regressionen

Series: Standardized Residuals	
Sample 1993Q2 2018Q4	
Observations 248751	
Mean	1.81e-17
Median	0.021895
Maximum	2.630740
Minimum	-2.284126
Std. Dev.	0.399323
Skewness	-0.337805
Kurtosis	4.237107
Jarque-Bera	20593.29
Probability	0.000000

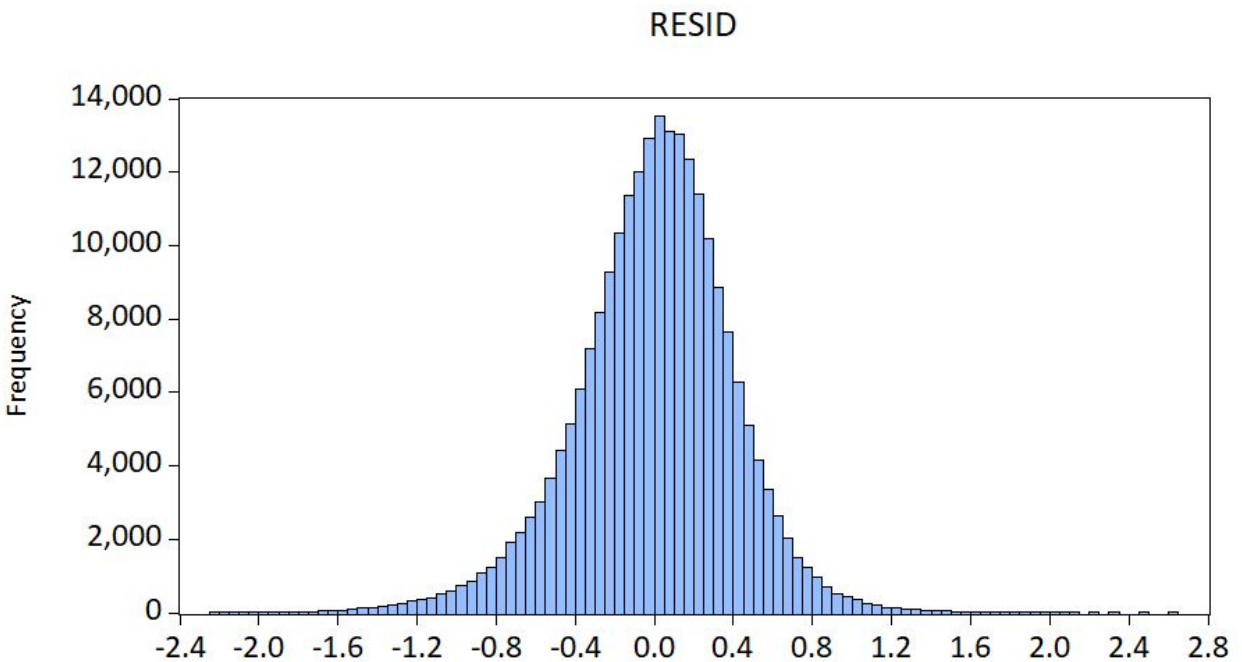
Figur 8.6: Deskriptiv statistik för feltermerna i den framtida regressionen

Series: Standardized Residuals	
Sample 1993Q4 2018Q4	
Observations 229649	
Mean	-6.12e-18
Median	0.020462
Maximum	2.629980
Minimum	-2.213528
Std. Dev.	0.395278
Skewness	-0.321662
Kurtosis	4.237996
Jarque-Bera	18625.50
Probability	0.000000

Figur 8.7: Histogram på feltermerna, regression med samtida mått på prestation



Figur 8.8: Histogram på feltermerna, regression med framtida mått på prestation



Figur 8.9: Resultat VIF test för regression med samtida prestationsmått

Variance Inflation Factors  
Date: 01/13/20 Time: 18:44  
Sample: 1993Q2 2018Q4  
Included observations: 248751

Variable	Coefficient Variance	Uncentered VIF	Centered VIF
SALES_GROWTH	2.70E-10	1.164324	1.022050
ROA	6.26E-08	1.069559	1.053059
RETURN	1.55E-09	1.035643	1.031435
LOG_MARKET_CAP	2.10E-06	539.6339	1.071644
C	0.000358	539.0522	NA

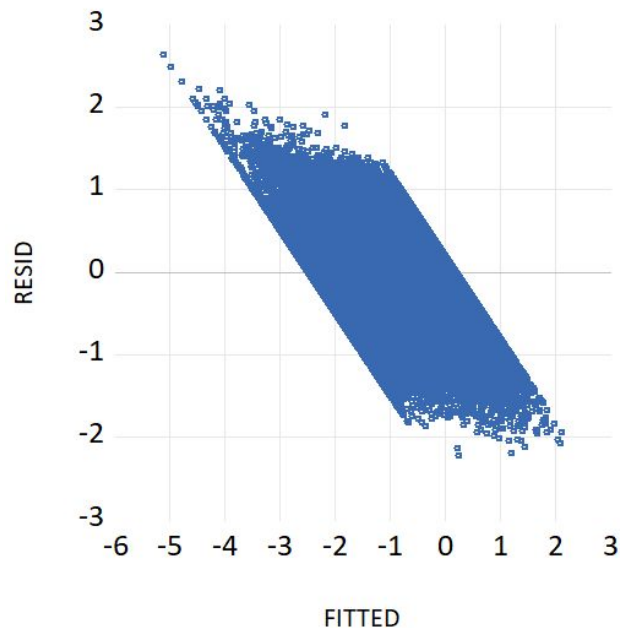


Figur 8.10: Resultat VIF test för regression med framtida prestationsmått

Variance Inflation Factors  
Date: 01/13/20 Time: 18:36  
Sample: 1993Q2 2018Q4  
Included observations: 229649

Variable	Coefficient Variance	Uncentered VIF	Centered VIF
SALES_GROWTH	3.13E-10	1.165349	1.022137
ROA	6.28E-08	1.071726	1.054455
RETURN	1.68E-09	1.039878	1.036303
LOG_MARKET_CAP	2.38E-06	580.6231	1.079203
C	0.000409	579.9714	NA

Figur 8.11: Scatter plot med residualer på y-axeln och estimerat tonläge på x-axeln



Figur 8.12: Hausmantest, cross-section random effects, regression med samtida mått på prestation

Correlated Random Effects - Hausman Test  
 Equation: EQ\_NOLAG  
 Test cross-section random effects

Test Summary	Chi-Sq. Statistic	Chi-Sq. d.f.	Prob.
Cross-section random	305.324636	4	0.0000

Figur 8.13: Hausmantest, period random effects, regression med framtida mått på prestation

Correlated Random Effects - Hausman Test  
 Equation: EQ\_NOLAG  
 Test period random effects

Test Summary	Chi-Sq. Statistic	Chi-Sq. d.f.	Prob.
Period random	94.244889	4	0.0000

Figur 8.14: Hausmantest, cross-section random effects, regression med framtida mått på prestation

Correlated Random Effects - Hausman Test  
 Equation: EQ\_LAG  
 Test cross-section random effects

Test Summary	Chi-Sq. Statistic	Chi-Sq. d.f.	Prob.
Cross-section random	330.470100	4	0.0000

Figur 8.15: Hausmantest, cross-section random effects, regression med samtida mått på prestation

Correlated Random Effects - Hausman Test  
Equation: EQ\_LAG  
Test period random effects

Test Summary	Chi-Sq. Statistic	Chi-Sq. d.f.	Prob.
Period random	88.085350	4	0.0000

Figur 8.16: Redundant fixed effects test, regression med samtida mått på prestation

Redundant Fixed Effects Tests  
Equation: EQ\_NOLAG  
Test cross-section and period fixed effects

Effects Test	Statistic	d.f.	Prob.
Cross-section F	20.939700 (8442,2402...		0.0000
Cross-section Chi-square	137197.597034	8442	0.0000
Period F	147.973618 (102,240202)		0.0000
Period Chi-square	15159.063514	102	0.0000
Cross-Section/Period F	22.537516 (8544,2402...		0.0000
Cross-Section/Period Chi-square	146441.932922	8544	0.0000

Figur 8.17: Redundant fixed effects test, regression med framtida mått på prestation

Redundant Fixed Effects Tests  
Equation: EQ\_LAG  
Test cross-section and period fixed effects

Effects Test	Statistic	d.f.	Prob.
Cross-section F	21.324897 (8023,2215...		0.0000
Cross-section Chi-square	131428.353352	8023	0.0000
Period F	123.788569 (100,221521)		0.0000
Period Chi-square	12487.317963	100	0.0000
Cross-Section/Period F	22.406954 (8123,2215...		0.0000
Cross-Section/Period Chi-square	137729.714748	8123	0.0000

## Bilaga 5: Pythonkod för tonlägesanalys

```
import Load_MasterDictionary as LM
import glob
import ntpath
from openpyxl import load_workbook
import re

MASTER_DICTIONARY_FILE = 'LoughranMcDonald_MasterDictionary_2018.csv'
NEGATIONS = ['NO', 'NOT', 'NONE', 'NEITHER', 'NEVER', 'NOBODY']
SENTIMENTS = ['negative', 'uncertainty', 'litigious', 'constraining', 'strong_modal',
'weak_modal'] #Positive ej inkluderad, hanteras separat

lm_dictionary = LM.load_masterdictionary(MASTER_DICTIONARY_FILE, True)

def main():
    print("main start")
    wbCount = str(4)
    wb_targets = load_workbook("output/targets.xlsx")
    ws_targets = wb_targets.get_sheet_by_name("targets")
    wb_results = load_workbook("output/results" + wbCount + ".xlsx")
    ws_results = wb_results.get_sheet_by_name("results")
    wb_wordcounts = load_workbook("output/wordcounts" + wbCount + ".xlsx")
    ws_wordcounts = wb_wordcounts.get_sheet_by_name("results")

    targets = dict()
    for r in list(range(2, 28)):
        y = str(int(ws_targets.cell(r, 1).value))
        targets[y] = dict()
        for c in list(range(2,14)):
            targets[y][str(c - 1)] = int(ws_targets.cell(r, c).value)

    print(targets)

    years = list(range(1993, 2019))
    months = list(range(1, 13))
    years.reverse()
    months.reverse()
    count = 0
    r = len(ws_results["A"])

    for y in years:
        yy = str(y)
        for m in months:
            if(targets[yy][str(m)] == 1): continue
            wDict = dict()
```

```

if(m < 10):
    mm = "0" + str(m)
else:
    mm = str(m)

print('starting:', yy, mm)

for fPath in glob.iglob('F:/data/' + yy + '/' + mm + '/*/*.txt', recursive=True):
    sDict = dict()
    sDict['positive'] = 0
    for s in SENTIMENTS:
        sDict[s] = 0
    count = count + 1
    wCount = 0
    negationCount = 4
    r = r + 1
    if r % 20 == 0:
        print("row:", r)
        print(fPath)

    txt = open(fPath, "r").read()
    tokens = re.findall('\w+', txt)
    for token in tokens:
        negationCount = negationCount + 1
        token = token.upper()

        if not token.isdigit() and len(token) > 1 and token in lm_dictionary:
            wCount = wCount + 1
            if token in wDict:
                wDict[token] += 1
            else:
                wDict[token] = 1

            if lm_dictionary[token].positive:
                if(negationCount <= 3):
                    sDict['negative'] += 1 #om negation word används i senaste tre orden, räkna positive
                    word som negative
                else:
                    sDict['positive'] += 1
            else:
                for s in SENTIMENTS:
                    if lm_dictionary[token].sentiment[s]:
                        sDict[s] += 1

    if token in NEGATIONS:
        negationCount = 0

```

```

#print neg, pos, etc. counts till excel
ws_results.cell(row=r, column=1).value = ntpath.basename(fPath)
ws_results.cell(row=r, column=2).value = wCount
ws_results.cell(row=r, column=3).value = sDict['positive']
c = 4
for s in SENTIMENTS:
    ws_results.cell(row=r, column=c).value = sDict[s]
    c = c + 1

```

```

#print wordcounts till excel
print("Printing wordcounts to Excel")
rr = 1
c = len(ws_wordcounts["1"])
ws_wordcounts.cell(row=1, column=c + 1).value = yy + mm
ws_wordcounts.cell(row=1, column=c + 2).value = "count"
for w in wDict:
    rr = rr + 1
    ws_wordcounts.cell(row=rr, column=c + 1).value = w
    ws_wordcounts.cell(row=rr, column=c + 2).value = wDict[w]

```

```

ws_targets.cell(row=(y - 1991), column=(m + 1)).value = 1 # markera target som klar

```

```

wb_targets.save(filename="output/targets.xlsx")
wb_results.save(filename="output/results" + wbCount + ".xlsx")
wb_wordcounts.save(filename="output/wordcounts" + wbCount + ".xlsx")

```

```

print("script end")

```

```

main()

```