Sentiment Analysis of 10-K Filings

An Approach to Automatic Processing of the Information
Hidden in Accounting Narratives

Copenhagen Business School
May 15, 2018
Cand.merc.(Fir)

Master's Thesis

Authors:
Jon Kandrup (971)
Johan Christian Mølhave (45092)

Supervisors:
Thomas Riise Johansen
Thomas Plenborg

96 Pages – 229.238 Characters (Including Spaces)

Abstract

In 10-K reports, there are numerous amounts of financials and accounting narratives available upon which investment decisions can be made. However, while financials relatively straightforward can be used to explain the performance of the firm, it is a more difficult task to use accounting narratives. Accounting narratives can contain information relevant to investors, such as expectations about the future and risk measures, which are not captured by the financials. However, given the scope of the 10-K report, it is a daunting task to find this information. This thesis seeks to provide a step in the direction of a more easy and automatic processing of the information hidden in the accounting narratives. The focus is on sentiment analysis since this provides a crude measure of whether the information contained in accounting narratives in the 10-K reports is favorable or unfavorable. The study is centered around a comparison of two sentiment analysis methods: The Bag of Words model and the Recursive Neural Tensor Network. In order to assess which model is superior in a financial setting, their extracted sentiment of the narratives in 10-K reports will be evaluated by their ability to explain stock returns. The approach of these models is to classify sentiment on a word and sentence level respectively, and they therefore represent a simple and a more sophisticated approach to textual analysis. The initial results showed that while the adjusted R^2 was remarkably low there was a statistically significant relationship between the models' sentiment score and the stock returns. However, after testing the validity of the results by adjusting the returns for systematic risk and including control variables, only the sentiment score of the Bag of Words model was significant in explaining the stock returns over the 10-K filing date. It is, therefore, concluded that further development of the Recursive Neural Tensor Network in order for it to be applicable to the financial domain is beneficial for the field of accounting research.

Table of Contents

Chapter 1 - Introduction

1.1 Motivation and Context

In financial markets, there are numerous amounts of quantitative and qualitative data available upon which investment decisions can be made. The literature in finance and accounting has predominantly been focused on the information content of quantitative measures when explaining stock behavior. The reason is that quantitative data is characterized by being easily available and seemingly more objective (Feldman, Givindaraj, Livnat, & Segal, 2009, p. 916). In addition, qualitative data is to some extent perceived as secondary, since it is used to explain the quantitative data. However, as Shiller (1981), Roll (1988), and Cutler et al (1989) demonstrates, implementing quantitative data alone to explain stock returns may be insufficient, and researchers have therefore looked elsewhere for additional explanatory variables. One area of research is to extract information from qualitative data such as the narrative in financial reports. The focus here is on accounting narratives produced by companies and aimed at shareholders. The narrative refers to words, such as stories and accounts, and is relevant for study because it plays a fundamental role in the way humans create subjective meaning (Beattie, 2014, p. 112). In addition, accounting narratives can contain information relevant to investors, such as expectations about the future and risk measures, which is not captured by the financials.

It is difficult, however, to find an objective quantitative measure of the qualitative information being conveyed in the accounting narratives, which makes it problematic to study the role and impact of these qualitative communications in the financial markets. In addition, the sheer amount of textual data makes it impossible for an investor to comprehend this information in order to make perfectly informed and rational decisions in a timely manner. One effect of this limitation is that when an annual report is published the textual information is not recognized in the share price immediately. The notion is discussed in more detail in Section 2.2.1.

However, given recent developments in computer science and linguistics, there are specific tools and models available in order to quantify the information content of qualitative data. Instead of investors having to read the textual part of an annual report, the linguistic model will process the qualitative data in a matter of seconds. Ideally, it will make it possible for the investor to meaningfully consider the qualitative information, thus optimizing the decision making, and reducing the amount of time needed to recognize the information in the textual parts of the annual report. In addition, this will help researchers in their work to understand the role and impact of accounting narratives in decision making.

There are various areas within the field of textual analysis such as targeted phrases, sentiment analysis, topic modeling and measures of documents similarity. This thesis will focus on sentiment analysis of financial reports and how it can be used to explain stock returns. The idea behind this relationship is, that if management shares truthful information in their narratives about prior and future performance of the firm that is not captured by the financials,

then market reactions should reflect the qualitative information disclosed by management. It is especially the sentiment of this narrative that is worthy of notice because it provides a crude measure of whether the performance is favorable or unfavorable (Feldman, Givindaraj, Livnat, & Segal, 2009, p. 951).

A lot of research has focused on the sentiment of narratives. The literature in finance and accounting has, however, primarily used a Bag of Words (BoW) approach to measure the sentiment of financial reports (this is covered in the literature review). This approach analyzes the text on a word level by counting the frequency of words, and can work well in some cases, however, from a linguistic point of view, ignoring word order when analyzing text is not sensible. One example of this is that companies have a tendency to frame a negative statement, by negating a sentence with many positive words (Loughran & McDonald, 2016, p. 1217). The BoW model will misclassify this sentence as positive because it will count the number of positive words over negative without considering the negation.

With this in mind, the purpose of this thesis is to present a new approach to sentiment analysis of financial reports by extending the analysis from word level to sentence level. This will be done by applying the Stanford CoreNLP framework, which is an open source Natural Language Processing software from the Stanford University Natural Language Processing Group (Stanford, 2018d). This software includes a sentiment classifier on a sentence level, referred to as a Recursive Neural Tensor Network (RNTN), which will be used to extract the sentiment of annual reports. It does this by breaking the sentences into meaningful components through deep parsing, thus, incorporating the information contained in the order of word sequences. Its results will be compared to the BoW approach to evaluate its use.

1.2 Research Question

The overall research question in this thesis is: To what degree can stock returns be explained by sentiment extracted from 10-K reports using the Stanford CoreNLP software and a Bag of Words approach.

To structure the answer to the research question, the following hypotheses will be tested:

H1: The Stanford CoreNLP[1] software can be used to explain stock returns by analyzing the sentiment of 10-K reports.

H2: The Stanford CoreNLP's sentiment analysis of 10-K reports is better at explaining stock returns than the Bag of Words approach using the Loughran & McDonald's (2011) financial dictionary to analyze the sentiment of 10-K reports.

These hypotheses are grounded in previous research as discussed in the Literature Review in Section 2.1.

[1] The Stanford Core NLP and Recursive Neural Tensor Network (RNTN) will be used interchangeably throughout the thesis.

1.3 Scope and Limitations

There are numerous ways and methods to answer the research question. It has therefore been necessary to make various limitations to remain within the scope of a master thesis.

Models for Textual Analysis

In the field of textual analysis, there are various models that can be used to answer the research question. This thesis will focus on the application and comparison of the following two models:

1. Bag of Words
2. The Stanford Core NLP sentiment classifier

The approach of these models is to classify sentiment on a word and sentence level respectively. The models, therefore, represent a 1) simple, and 2) a more sophisticated approach to textual analysis. Rather than evaluating the math behind the models, the focus will be on the application of them, and their ability to explain movements in stock price, when they are applied to 10-K reports.

Data:

A lot of different qualitative financial data can be used as input in the models. It is, therefore, necessary to limit the amount of data that is used. In this thesis, the focus will be on 10-K filings from firms in the S&P 500 index from 2008 to 2017. The reason behind this choice will be described in section 2.4 and 5.2.

Labelled Corpora

When using RNTN and BoW models labelled corpora are needed. In the context of the Natural Language Processing field labelled corpora are dictionaries of either words or phrases that have been assigned a specific value such as positive or negative. Since the process of making our own corpora requires an extensive amount of work, which is beyond the scope of this thesis, we have chosen to use corpora that have been created in previous research. For the BoW model there is a corpus that made by Loughran and McDonald (2011) specifically for financial text, however, the RNTN does not have that option. Even though the RNTN can be trained on a new corpus it must be in a specific format called a sentiment treebank, which Loughran and McDonald's corpus does not match. Unfortunately, there are no publicly available sentiment treebanks made specifically for financial texts. In this paper, the expected implications of this shortcoming are that it will have a negative influence on the RNTN's output.

1.4 Structure of the Thesis

The thesis is structured into 8 chapters. The structure of these chapters will be described below.

Chapter 1 - Introduction

The introduction to the master thesis is given in chapter 1, which is also the current chapter. This introduction consists of the motivation and context, the research question, limitations and scope as well as an overview of this thesis.

Chapter 2 - Background:

This chapter starts with a discussion of, whether there is theoretical evidence to believe that textual analysis can detect patterns in the stock market. Thereafter follows a brief description of 10-K reports and the S&P500 and the benefits of using them for the research in this thesis. Lastly, a review of the academic literature regarding textual analysis, its general development and existing use in finance, will be presented.

Chapter 3 – Model Description:

The purpose of this chapter is to uncover the theory behind the BoW approach and RNTN used in this thesis. First, the theory of content analysis is introduced. Thereafter, machine learning is introduced with a description of one of its most principal models: The Neural Network. Lastly, some key methods of natural language processing are introduced along with a description of the RNTN and a discussion of its applicability.

Chapter 4 – Extraction of Sentiment

This chapter will explain how the 10-K narratives will be transformed into a quantitative sentiment score by the programming behind the BoW model and the software the RNTN uses. In addition, an explanation is given of how these classifications have been transformed into independent variables of concern. Lastly, a discussion of the pitfalls there might be when interpreting the BoW's and RNTN's sentiment scores in the further analysis is presented.

Chapter 5 – Data and Regressions

The aim of this chapter is to discuss and explain the choices of data. The first section describes the different data sources that have been used, the selection of appropriate data for analysis purposes, and how this data has been retrieved. The second section addresses the challenges of parsing the 10-K reports. Finally, the analysis is in focus, where the methodological considerations behind the regression between the sentiment scores and the stock returns are presented.

Chapter 6 – Results:

This chapter contains the results of the different regressions that have been performed. The results are analyzed, where the aim is to examine if the sentiment scores of the BoW and RNTN are able to capture the favorable or unfavorable information in the narratives of 10-K reports, and by doing that predict the changes in stock returns over the filing date. To ensure the robustness of the results, different regressions and tests will furthermore be performed in this chapter. Finally, the conclusions of the analysis will be used to answer the research question and hypotheses of this thesis.

Chapter 7 – Discussion of Results:

This chapter provides a discussion of the results and insights of this thesis, thus, giving a perspective of how to interpret the results, its limitations, and where and how to direct efforts in future research. Firstly, a discussion of factors that may influence the results and the interpretation of them is presented. Lastly, the contribution of this thesis and suggestions for future research prospects is discussed.

Chapter 8 - Conclusion:

This chapter concludes the findings in this thesis and provides a perspective on its contributions.

Chapter 2 – Background

This chapter describes the preliminary study carried out to establish a foundation that can be used to support the methodical considerations.

First, a review of the academic literature regarding textual analysis, its general development, and its existing use in finance will be presented, which will form the basis for the Hypotheses. Second, a discussion of whether there is reason to believe that textual analysis can detect patterns in the stock market will be presented, which is followed by a description of the artificial intelligence field, which the models used in this thesis relate to. Last, the 10-K reports that are the field of research are described.

2.1 Literature Review and Hypotheses

The following section reviews the academic literature regarding textual analysis and its use in accounting. The review also covers methods of textual analysis recently developed in computer science. The most noteworthy research in this area will be covered, however, the concepts that are deemed necessary to understand the RNTN model will receive the most elaboration. Lastly, the two areas are compared to reveal uncovered areas of research in the literature, which will be used to form the hypotheses of this thesis.

2.1.1 Textual Analysis of Financial Report Narratives

The financial reporting environment is complex. There are many parties involved in the information production such as preparers, auditors, and the media and the behavior of these parties is similarly complex. Even though the reports are standardized, their information output and its interpretation reflect the complexity of the environment of which they are created. This makes the information extraction of these reports applicable to many different fields of research, some of which will be reviewed in this section with an emphasis on the methods of extracting information from accounting narratives.

The research on this information is related to two areas: the literature on accounting narratives and that of voluntary disclosure. Disclosure research draws upon economic information asymmetry arguments and agency theory, where disclosed information is viewed as a rational trade-off between costs and benefits. Information asymmetry reduction is the benefit of extensive disclosure as it leads to a reduction in the cost of capital and increased share price and liquidity. This comes at various economic costs such as the loss of competitive advantage since secrets about the sensitive information about the business model might be revealed (Beattie, 2014, p. 112)

There has, however, been a "turn" of interest towards the narrative of these financial disclosures, where narrative refers to the words and stories management uses. The "narrative turn" refers to the interest in the narrative in literary studies that spread to many other scientific disciplines such as accounting. This interest was sparked by the recognition in the 1980's humanities and social sciences that narrative plays a fundamental role in the way humans

create subjective meaning. Research into accounting narratives broadly covers a spectrum from large-scale quantitative analysis with roots in economic theory (Li, 2008) and social sciences (Merkl-Davies, Brennan, & McLeay, 2011) to qualitative case studies using methods from the humanities (Davidson, 2008) (Beattie, 2014, p. 112).

Soper and Dolphin (1964) is one of the earliest papers on accounting narratives and was published over fifty years ago. This paper discussed the research of readability in relation to the understandability of financial reports. In Adelberg's (1979) paper the first mention of the term "narrative" in relation to accounting disclosures appeared. The paper explained the frictions between the two roles of accounting narratives – communication or manipulation. The ideas of manipulation and especially impression management are effects of the entering of psychology and social psychology into accounting research. Impression management, in the context of textual and visual aspects of financial reporting, can be viewed as embodying the literature of the (earnings) management of accounting numbers. A large part of the textual studies in this area was content oriented, focusing mainly on keywords such as positive/negative to explain the content of a text (Beattie, 2014, p. 114). From the 1990s onward, content analysis became a commonly used research method in accounting, where the dominant method was to transform the text into category numbers that expressed a summary description of the text. Content analysis was applicable to a number of studies such as the link between company performance and the ratio of positive/negative keywords as a proxy for narrative tone (e.g. Clatworthy & Jones (2003)). Because of the limitations of computer technology, this type of analysis was done manually at that time (Beattie, 2014, p. 114).

Around the new millennium, there was a recognition in the literature of accounting that there was a need for methods innovation in relation to the developments in computer science in order to conduct large-scale studies (Core, 2001). A notable response to this recognition was Tetlock (2007), Li (2008), and Ronen Feldman (2009), who examines links between linguistic sentiment, readability and market returns by applying computer science to content analysis. The renewed interest around this time can be explained by (i) the growing availability of digitized text, (ii) the development of increasingly sophisticated computerized software which permits large-sample studies, and (iii) a concern with finding ways to enhance the predictive value of financial reporting due to the observed decline in the value relevance of financial statements (Francis & Schipper (1999)) (Beattie, 2014, p. 116).

The three factors mentioned above inspired further research of the links between linguistic sentiment and financial performance. Notable research includes Ferris et al (2013) who by analyzing Initial Public Offerings (IPO) prospects found that prospectus conservatism using a negative sentiment score based on the Loughran-McDonald dictionary (2011)) in IPO prospects is positively related to underpricing. Complementing Ferris et al's findings, Brau et al (2016) finds that more frequent use of positive and/or less frequent use of negative strategic words in IPO documents leads to more IPO underpricing. Regarding the tone of financial reports, Yekine et al (2016) and

Jegadeesh (2013) find the positiveness and negativity of financial reports has an effect on the market reaction, and Lou et al (2017) find that this relationship is more pronounced with positive tones in the earnings announcements issued by companies with more competent management teams. In addition, apart from financial reports, news sources have been data mined for information about financial performance. Examples of this kind of research are Ahmad et al (2016), Das and Chen (2007), and Tsai et al (2016) who reveal that sentiment analysis of news has relationships with stock returns and credit risk evaluations respectively.

2.1.1.1 Common Versus Financial Dictionaries for Word Counting

A relevant finding is Li (2010) who uses the simple machine learning algorithm: The Naïve Bayesian machine learning algorithm to examine the information content of the forward-looking statements in the Management Discussion and Analysis section of 10-K and 10-Q filings. He measures the tone based on three commonly used dictionaries for word counting (Diction, General Inquirer, and the Linguistic Inquiry and Word Count), and find that they do not positively predict future performance. He suggests that these dictionaries might not work well for analyzing corporate filings. This, however, does not invalidate the results from research based upon these dictionaries, such as Davis et al (2006), who uses Diction to find that there is a positive (negative) association between optimistic (pessimistic) language usage and future firm performance, and Kothari et al (2009), who uses the General Inquirer to find negative disclosures from business press sources result in increased cost of capital and return volatility, and favorable reports reduce the cost of capital and return volatility.

Tim Loughran and Bill McDonald have contributed with much research into word counting in accounting and have some of the most cited research papers in this area. Their work builds upon Li's (2010) findings showing that word lists developed for other disciplines misclassify common words in financial texts, and develop an alternative negative word list, along with five other word lists, that better reflect tone in financial texts (Loughran & McDonald, 2011). Complementing this work, they also argue that Diction is inappropriate for gauging the tone of financial disclosures, and The Loughran-McDonald dictionary (Loughran & McDonald, 2011) appears better at capturing tone in business text than Diction (Loughran & McDonald, 2015). Since making this dictionary for financial reports, it has become widely used for research regarding sentiment analysis of financial text (Loughran & McDonald, 2016, p. 1206).

Besides making a dictionary for financial texts, they find evidence that phrases like unbilled receivables signal a firm may subsequently be accused of fraud. At the 10-K filing date, phrases like substantial doubt are linked with significantly lower filing date excess stock returns, higher volatility, and greater analyst earnings forecast dispersion (McDonald & Loughran, 2011). In addition, they find that IPOs with high levels of uncertain text have higher first-day returns, absolute offer price revisions, and subsequent volatility (Loughran & McDonald, 2012). In 2015 they created a measure for the inherent trust in a company by counting the number of times 21 trust-related words appear in the Management Discussion & Analysis section of the annual report. They find that firms who score high on their

trust-proxy frequently use audit- and control-type words and the trust-proxy is positively linked with subsequent share price volatility (Audi, Loughran, & McDonald, 2015).

2.1.2 Deep Learning for Opinion Mining

This section will describe the literature of deep learning for textual analysis recently developed in computer science.

Deep learning is an approach with multiple levels of representation learning, which has become popular in applications of computer vision, speech recognition, and natural language processing. In this section, there will be introduced some successful deep learning algorithms for natural language processing. With the rapid growth of deep learning, many recent studies expect to build vectors as text features for opinion mining without any need for manual feature learning that requires labeled data. Currently, however, the task of opinion expression extraction is formulated as a token-level pattern recognition which involves the assignment of a categorical label to each member of a sequence of values (assigning grammar or sentiment to each word in a sentence). In order to address this, a lot of studies use Conditional Random Field (CRF) or semi-CRF with manually designed discrete features such as word features, phrase features, and syntactic features in order to identify opinion expressions and the sources of the opinions, emotions, and sentiments (Cardie, Choi, & Breck, 2007).

2.1.2.1 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are popular models that have shown great promise in many NLP tasks. An RNN is an extension of a conventional neural network, which is able to handle variable length input sequences. Thus, RNNs are naturally applicable for language modeling and other related tasks. Irsoy and Cardie (2014) applied Deep Recurrent Neural Networks (DRNNs) to extract opinion expressions from sentences and showed that DRNNs outperform CRFs. The method is constructed by stacking multiple layers of RNNs on top of each other. Every layer of the DRNN treats the memory sequence from the previous layer as the input sequence and computes its own memory representation, thus bringing a temporal hierarchy to the architecture.

Over the years researchers have given much attention to the enhancement of the RNNs. This has amongst others resulted in the development of the Bidirectional RNNs, which are based on the idea that the output at time t may depend on not only previous elements in the sequence but also future elements. Bidirectional RNNs are quite simple in the sense that they are two RNNs stacked on top of each other. The output is then computed based on the hidden states of both RNNs. Despite its simplicity, it is a powerful tool in NLP given that it is able to predict a missing word in a sequence by looking at both the left and the right context. A natural augmentation to this model is the deep bidirectional RNNs, which operates with multiple layers per time step that in practice this gives a deeper learning capacity (Sun, Luo, & Chen, 2017, p. 20).

2.1.2.2 Semantic Vector Spaces

In computer science, semantic vector spaces for single words are representations of the meaning of these words and have been widely used as features (Turney & Pantel, 2010). However, because they cannot capture the meaning of longer phrases properly, compositionality in semantic vector spaces received a lot of attention by (Mitchell & Lapata, 2010), (Socher, Manning, & Ng, 2010), (Zanzotto, Fallucchi, Korkontzelos, & Manandhar, 2010), (Yessenalina & Cardie, 2011), (Socher, Manning, & Huval, 2012), and (Grefenstette, Dino, Zhang, Sadrzadeh, & Baroni, 2013). This research was held back by the lack of labeled compositionality resources, such as sentiment treebanks. Such a resource would ideally make it possible to train models upon that would be able to reflect the meaning of phrases and sentences instead of only words. Therefore, Socher et al. (2013) proposed a model called Recursive Neural Tensor Network (RNTN) for sentiment analysis together with The Stanford Sentiment Treebank, which was the first of its kind. They represented a phrase through word vectors and a parsing tree and then computed the vectors for higher nodes in the tree through the same tensor-based composition function. The RNTN model can capture the effects of negation and its scope at various tree levels for both positive and negative phrases. This is the same model that this thesis uses for sentiment extraction of 10-K reports.

2.1.2.3 Long Short-Term Memory

Long Short-Term Memory (LSTM) (Schmidhuber & Hochreiter, 1997) is specifically designed to model long-term dependencies in RNNs. LSTMs do not have a fundamentally different architecture from RNNs, but they use a different function to compute the hidden states. The memory function in LSTMs are called cells which take the previous state h_{t-1} and current observation x_t as inputs. Internally these cells decide what to keep in and what to erase from memory. They then combine the previous state, current memory, and current observation. It turns out that these types of units are very efficient at capturing long-term dependencies, which makes the LSTM very applicable to natural language. Sequential models like RNNs and LSTMs are also verified as powerful approaches for semantic composition in the same sense as the RNTN model (Tai, Socher, & Manning, 2015). Liu et al. (2015) proposed a general class of discriminative models based on pre-trained RNNs and word embeddings that can be successfully applied to fine-grained opinion mining without any task-specific feature engineering effort.

2.1.2.4 Convolutional Neural Networks

Another powerful neural network for sentence representation is Convolutional Neural Networks (CNNs). Kalchbrenner el at. (2014) described a convolutional architecture called Dynamic Convolutional Neural Networks (DCNNs) for semantically modeling of sentences. The network uses dynamic k-max pooling, a global pooling operation over linear sequences. The network handles input sentences with variable lengths and induces a feature graph over the sentences that is capable of capturing short and long-range relations.

2.1.2.5 Word Representation and Embeddings

Meanwhile, the advances in word representation and embeddings using neural networks have contributed to the advances in opinion mining by using deep learning methods (Sun, Luo, & Chen, 2017, p. 20). A pioneering work in this field is given by Bengio et al. (2003). The authors introduced a neural probabilistic language model that learns a continuous representation for words and a probability function for word sequences based on the word representations. Mikolov et al. (2013) and Mikolov et al. (2013b) introduced Continuous Bag-of-Words (CBOW) and skip-gram language models and released the popular word2vec 10 toolkit. The CBOW model predicts the current word based on the embeddings of its context words, and the skip-gram model predicts surrounding words according to the embedding of the current word. The word2vec 10 toolkit provides an easy method for constructing these vectors which fit as input into NLP algorithms. Pennington et al. (2014) introduced Global Vectors for Word Representation (GloVe), an unsupervised learning algorithm for obtaining vector representations of words. Training is performed on aggregated global word-word cooccurrence statistics from a corpus, and the resultant representations show interesting linear substructures of the word vector space.

2.1.3 Hypotheses

As Loughran and McDonald concluded in their 2016 survey (Loughran & McDonald, 2016, p. 1223) much of the literature in finance and accounting uses a Bag of Words approach to measure document sentiment of financial reports. Thus, despite the, as mentioned before, abundant research into deep learning for opinion mining, there is hardly any research into whether it is applicable to financial reports, which is why according to Loughran and McDonald this is clearly an area for future research. Specifically, the question still remains unanswered as to whether there is meaningful information to be obtained in financial reports by breaking down the sentences into meaningful components through deep parsing and consequently incorporating the information contained in the order of word sequences (Loughran & McDonald, 2016, p. 1223).

One way to explore this area is conducting a study of stock returns by analyzing 10-K reports with the Stanford CoreNLP framework, which is an easily accessible open software that uses deep learning for opinion mining. One drawback of the software is that the sentiment classifier has been trained on the Stanford Sentiment Treebank, which encompasses a language used in movie reviews. Li (2010) and McDonald et al (2011) found that conventional dictionaries might not be good at capturing the information in financial disclosures. There are, however, no sentiment treebanks based upon the language used in finance (Kearney & Liu, 2014, p. 177) (Beattie, 2014, p. 128). In addition, results from existing research based on common dictionaries (such as Kothari et al (2009) and Davis et al (2006)) still hold, suggesting that it is possible to find results despite using a relatively noisy dictionary. In addition, Vivien Beattie advocates for the use of mixed methods and theoretical pluralism in the research into accounting narratives (Beattie, 2014, p. 128). Therefore, there is innovation and insights to be gained by mixing methods of

different scientific fields in the accounting research. One example of this would be the use of the theories of deep learning from computer science. Thus, as presented in Section 1.2, the first hypothesis in the thesis is the following:

H1: The Stanford CoreNLP software can be used to predict stock returns by analyzing the sentiment of 10-K reports.

In order to assess the significance of the results of **H1** a contrast between the advanced NLP method that classifies sentiment on a sentence level and the simpler method that classifies sentiment on a word level, but is characterized by economic theory, will be beneficial. This will be done by comparing the results of the Stanford CoreNLP with a Bag of Words approach using the Loughran & McDonald's (2011) financial dictionary to evaluate its use. Hence, the second hypothesis is:

H2: The Stanford CoreNLP's sentiment analysis of 10-K reports is a better predictor of stock returns than the Bag of Words approach using the Loughran & McDonald's (2011) financial dictionary to analyze the sentiment of 10-K reports.

2.2 Market-Based Accounting Research

In market-based accounting research, the purpose is to examine the relationship between publicly disclosed accounting information, and the consequences of the use of this information by equity investors. The effect of these different disclosures is reflected in the movements in stock prices of stocks traded in different exchanges. The general assumptions in market-based accounting research are the existence of efficient capital markets and the validity of the Capital Asset Pricing Model (CAPM) theory (Lev & Ohlson, 1982, p. 249+283).

2.2.1 Efficiently Inefficient Markets

The efficiency of financial markets is described by the efficient market hypothesis. Pedersen (2015) describes that the spectrum starts from fully efficient markets, where the idea is that all prices reflect all relevant information at all times. The other end of the spectrum is the inefficient market, where market prices are believed to be significantly influenced by investor irrationality and generally have little relation to firm fundamentals due to naïve investors. In early studies, the belief was that the market was in between these extremes on a semi-strong level, which reflects that all publicly available information is incorporated in the price, when information is published (Lev & Ohlson, 1982, p. 284).

In later research, this belief has been modified since it has been shown that there is a significant post-earnings announcement drift. This post-earnings announcement drift has shown to be present up to 60 trading days after the earnings announcement, but it is most notable in the first 5 trading days (Bernard & Thomas, 1989, p. 11+13). This is a clear indication that new information is incorporated in the price in short matter of time, but not to a full extent which should be taken into account.

It can, therefore, be concluded that markets are somewhat efficient given that the textual information is reflected in the share price, however, they are not completely efficient since there is a time-lag before the new information is incorporated in the share price. Thus, it seems that the market is more in line with Pedersen's (2015) theory of efficiently inefficient markets. This is the idea that markets are inefficient but to an efficient extent, where competition among professional investors makes markets almost efficient, however, the market remains so inefficient that they are compensated for their costs and risks.

Based on this, the expectation is that the potential correlation between the share price and sentiment score will be affected by the time window between the date of the annual report and the inclusion of new information in the share prices.

2.2.2 The Validity of the CAPM

In market-based accounting research, the CAPM theory is assumed to be valid and has therefore often been used to adjust the stock returns of systematic risk. It has, however, been questioned whether it is appropriate to use the CAPM as the only measure of adjustment (Lev & Ohlson, 1982, p. 283+287). One contribution to this issue came from Eugene Fama and Kenneth French (1992) who proposed a three-factor model, which on top of the market index takes firm size and a book-to-market ratio into account. The inclusion of these market factors is empirically motivated since it has been shown that historical average returns on stocks of small firms and stocks with high ratios of book equity to market equity are higher than predicted by the security market line of the CAPM (Bodie, Kane, & Marcus, 2014, p. 426). To adjust for the inherent systematic risk in stock prices the Fama-French three-factor model will be used, which is given by the following formula:

$$E(r_i) - r_f = \alpha_i + \beta_i * R_M + s_i * SMB + h_i * HML$$

Where the coefficients $\beta_i, s_i \; and \; h_i$ are betas (loading) of the stock on the three factors, where β_i is the loading of the market index, s_i is the loading of the firm size variable and h_i is the loading of the book-to-market ratio variable (Bodie, Kane, & Marcus, 2014, pp. 427-428). The three variables SMB, HML, and R_M are the different returns, which the factor loadings are multiplied by. The SMB is the average return of a portfolio of small stocks in excess of the return on a portfolio of large stocks. The HML variable is the average return of stocks with a high book-to-market ratio in excess of the return on a portfolio of stocks with a low market-to-book ratio. Finally, the R_m variable is the excess return on the market. Adjusting for these factors will leave a return that has been "cleaned" from factors that explain around 90% of the diversified return (Fama & French, 1992). Regressing this return on the sentiment score of the models will yield a more accurate relationship between the sentiment of the narratives in the 10-K report and the following stock return.

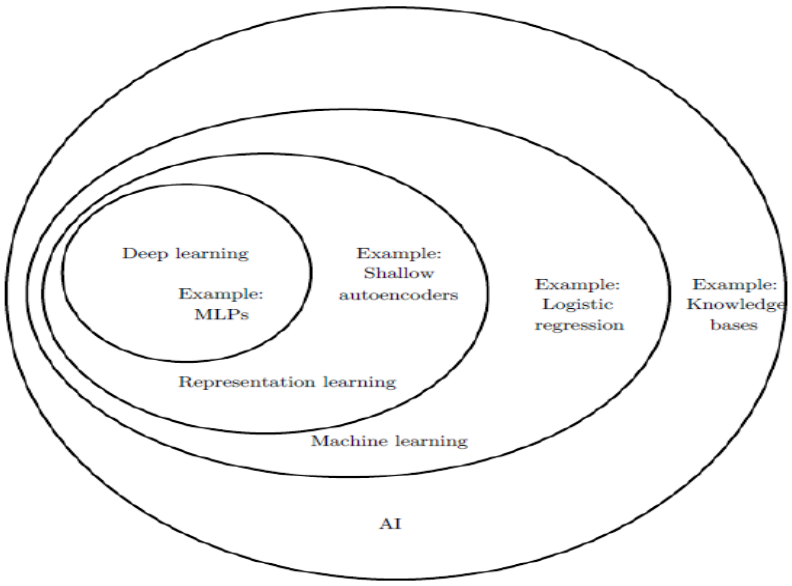
2.3 Artificial Intelligence, Machine Learning, and Deep Learning

This section provides an overview of the scientific disciplines which the two models in this thesis are part of.

The idea behind computer-based Artificial Intelligence (AI) dates back to 1950 when Alan Turing proposed the Turing test, which is: "can a computer communicate well enough to persuade a human that it, too, is human?" (McKinsey&Company, 2017, p. 9).

In the field of AI there are different disciplines. Their relationships are illustrated in the following Venn diagram:

Figure 2.1: Venn diagram of Artificial Intelligence



Source: (Goodfellow, Bengio, & Courville, 2016, p. 9)

The models used in this thesis come from different disciplines in the AI field. The BoW model is equivalent to a knowledge base since it extracts information or knowledge from its sentiment dictionary. It is, therefore, part of the broader AI field. The Stanford Core NLP sentiment classifier, on the other hand, uses a model that belongs to the deep learning discipline.

Deep learning is a part of machine learning, which is concerned with the challenge of constructing computer programs that automatically improve with experience. There are two definitions of machine learning. In 1959 Arthur Samuel, who was a pioneer in the field of artificial intelligence, coined the term "Machine Learning" describing it as:

" the field of study that gives computers the ability to learn without being explicitly programmed" (Samuel, 1959).

Tom Mitchell provides a more modern definition:

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E" (Mitchell T. M., 1997).

Since the first ideas of Artificial Intelligence were proposed in the 1950's a lot of progress has been made in the machine learning field. This development has especially been accelerated in the 21st century. The reason is that the different Machine Learning models are now able to be trained on a sufficient amount of data, which is made possible due to faster computers (McKinsey&Company, 2017, pp. 6-9).

What separates deep learning from regular machine learning is that deep learning can be regarded as models that either involve a greater amount of composition of learned functions or learned concepts than traditional machine learning does (Goodfellow, Bengio, & Courville, 2016, p. 8). A typical example of a deep learning model is the neural network, which is essentially layers of stacked sigmoid (learning) functions. This allows representations of the world as a nested hierarchy of concepts that all have a relation to each other.

These properties of deep learning make it highly applicable to Natural Language Processing given the characteristics of language. Sentences are largely defined by being a sequence of inputs that have different relations to each other across time. One word such as "not" in a sentence might have a profound impact on the semantic meaning of the rest of the words in the sentence. This is difficult for conventional statistical models to capture, however, a neural network with the correct architecture has better prerequisites for this. The Stanford CoreNLP software used in this thesis utilizes the Recursive Neural Tensor Network which is a subset of a regular neural network. This model, together with the core principles of a regular neural network will be explained in more detail in Section 3.2.

2.4 10-K Reports

This section will give an overview of 10-K reports. The thesis seeks to attain knowledge about the accounting narratives that contain truthful information about prior and future performance of the firm that is not captured by the financials. These narratives can be found in the 10-K reports, which is why a description of the 10-Ks will be given.

In the USA the federal securities laws require three different types of companies to file annual reports with the U.S. Securities and Exchange Commission (SEC) on an ongoing basis. These types of companies are:

1. A company having a class of security listed on a national securities exchange.
2. Unlisted companies with more than $10 million of assets and more than 2.000 security holders.
3. Companies registering either equity or debt securities under the Securities Act.

(EY, 2017, p. 5)

The annual reports filed with The U.S. Securities and Exchange Commission (SEC) are called a 10-K report which is a standardized format that the SEC requires companies to submit their annual reports in (U.S. Securities and Exchange Commission, 2009b). When the 10-K reports are filed with the SEC, they are gathered in a database called EDGAR, and through this made easily available for the public to download (Loughran & McDonald, 2017, p. 1). Even though companies have filed a 10-K report to the SEC, they will often make an annual report for the investors as well. The 10-K report and annual report are similar in many ways, but there are some differences (U.S. Securities and Exchange Commission, 2014).

The annual report is presented in a more professional and marketable way since its intended recipients are the shareholders of the company. The 10-K report, on the other hand, is not designed with investors in mind and is therefore often longer and harder to process than annual reports. The 10-K report gives a full description of the company's financial activity during the previous fiscal year. Furthermore, the information that must be included and how it should be organized is highly regulated by the SEC. Examples of what should be included is a detailed picture of the company's business, the risks it faces, the companies operating and financial results for the past fiscal year, and finally, the management's perspective of the financial conditions and results in a narrative form, which is of particular interest in this thesis (U.S. Securities and Exchange Commission, 2011).

The 10-K report is organized in 4 parts, which is shown in table 2.1, where part 1 describes different company-specific information, part 2 describes how the company has performed in the previous year and how their outlook for the future is, part 3 describes different corporate governance issues and finally part 4 which consists of different exhibits. Table 2.1, furthermore, shows the full list of items that the 10-K report includes, and how they are organized.

The 10-K report includes 20 different items, where the most interesting narratives are found in item 7, which is the "Management's Discussion and Analysis of Financial Condition and Results of Operations"(MD&A). In this item, management is required in a narrative form to comment on the company's current financial conditions, changes in these financial conditions, results of operations, and an outlook about the future as well (EY, 2017, pp. 71-72). A further description of the MD&A is given in Section 5.1.2.3.

The 10-K report is viewed as a good choice to base the thesis on because the 10-K reports are made available and are easy to download at the EDGAR database. The 10-K is furthermore highly regulated by SEC regarding the information that should be included, which makes it easier to compare the different companies. The last important factor is that management is required to give their perspective of the financial conditions and results in a narrative form, which provides interesting input for automatic textual analysis. It can, however, be discussed if the annual report prepared for investors might have been a better choice, but since the information in the two types of reports are identical in most cases, the 10-K reports are easier to download, they are highly regulated by the SEC, and more standardized, the 10-K reports are viewed as the better choice.

Table 2.1: 10-K report

Item	Heading
	Part 1
Item 1	Business
Item 1A	Risk Factors
Item 1B	Unresolved Staff Comments
Item 2	Properties
Item 3	Legal Proceedings
Item 4	Mine Safety Disclosures
	Part 2
Item 5	Market for Registrant's Common Equity, Related Stockholder Matters and Issuer Purchases of Equity Securities
Item 6	Selected Financial Data
Item 7	Management's Discussion and Analysis of Financial Condition and Results of Operations
Item 7A	Quantitative and Qualitative Disclosures about Market Risk
Item 8	Financial Statements and Supplementary Data
Item 9	Changes in and Disagreements with Accountants on Accounting and Financial Disclosure
Item 9A	Controls and Procedures
Item 9B	Other Information
	Part 3
Item 10	Directors, Executive Officers, and Corporate Governance
Item 11	Executive Compensation
Item 12	Security Ownership of Certain Beneficial Owners and Management and Related Stockholder Matters
Item 13	Certain Relationships and Related Transactions, and Director Independence
Item 14	Principal Accountant Fees and Services
	Part 4
Item 15	Exhibits, Financial Statement Schedules

Source: (U.S. Securities and Exchange Commission, 2011)

Chapter 3 – Model Description

The purpose of this chapter is to uncover the theory behind the BoW approach and RNTN used in this thesis. First, the theory of content analysis is introduced. Thereafter, machine learning is introduced with a description of one of its most principal models: The Neural Network. Lastly, some key methods of natural language processing are introduced along with a description of the RNTN and a discussion of its applicability.

3.1 Content Analysis and Bag of Words (BoW)

Content analysis is the primary scientific tool of this thesis. Krippendorff (2004, p. 18) defines content analysis as: *"... a research technique for making replicable and valid inferences from texts (or other meaningful matter) to the contexts of their use"*. Content analysis has a broad area of different techniques, in this thesis, the dictionary approach will be used. The theory of semantics (meaning) that dominates the dictionary approach is derived from taxonomy by assigning classifiers to text. The idea is that texts can be represented on different levels of abstraction (such as classifying a text as overall positive or negative) and the meanings are distributed in a body of text and need to be identified and extracted (Krippendorff, 2004, p. 283). The dominant way of doing this is through obtaining frequencies, not of the actual characters in the text, but of word families that share the same meaning. Words are distributed to different word families that share the same meaning such as positivity, negativity, uncertainty and more. The word families are part of an overall dictionary typically incorporating some theme, such as Loughran and McDonald's financial dictionary for sentiment analysis (Loughran & McDonald, 2011). Thus, there are different dictionaries, and the choice of dictionary has a large say in what content will be extracted.

After obtaining the frequencies of the word families that are defined by the dictionary, each word family is compared to each other to infer meaning from the overall text. For example: if the word family of positivity is larger than that of negativity, one can infer that the text is more positive than negative. This approach is also called the Bag of Words approach because the summarization of words into word families can be seen as having various "bags" of words, where the size of these bags is used to infer meaning. This approach is regarded as fairly simple, however, because it makes an assumption of independence between words meaning that the order of the sequence of words and their context is not important (Loughran & McDonald, 2016, p. 1199).

The RNTN is similar to the BoW approach in that it utilizes a dictionary (sentiment treebank) to obtain frequencies of meaning. These frequencies are then used to make inferences from texts. This makes it also a part of the content analysis field in that sense. The difference is, however, that except the frequencies being on a word level, it is on a sentence level.

This approach to content analysis of 10-K reports is rather crude. However, since it is made on a large scale quantitative level, it will be possible to capture if there is an overall trend between the sentiment of 10-K reports and stock returns. The practical approach of the content analysis applied in this thesis will be further elaborated in Chapter 4.

3.2 The Recursive Neural Tensor Network (RNTN)

This section will describe the model that is used for classifying the sentiment of 10-K narratives on a sentence level. In order to better understand the RNTN, an introduction to neural networks and semantic representation of words is presented. Thereafter follows a description of the intuition behind the RNTN, and the reason for choosing it for this thesis. Finally, a step by step notation of the model is presented.

3.2.1 Neural Networks

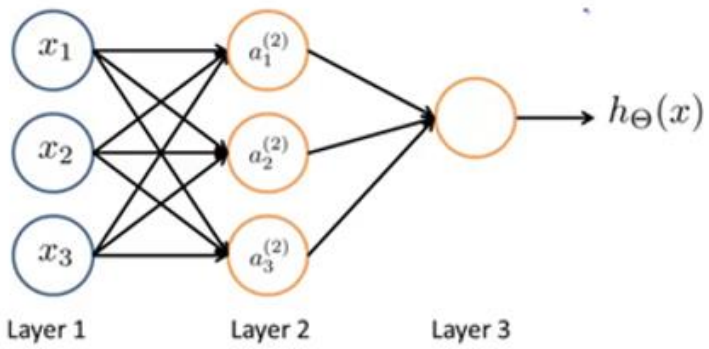
In order to better understand the RNTN, an overall explanation of a simple neural network and how it "learns" from data is presented.

An artificial neural network is a system that is inspired by the biological neural network in brains. Instead of using code to perform a specific task, the artificial neural network learns to perform those tasks simply by observing examples and adjusting its parameters until it can replicate the results. One example could be that the neural network predicts house prices from some input variables, x, that could be the size, the number of bedrooms, the zip code, and how wealthy the municipality is. The advantage of a neural network is that it would be able to find some additional attributes on its own (such as the quality of the schools, the walkability of the area, and family sizes) which explains the relationship between the price and the house. Its ability to do this depends on the type of neural network and the data you feed it. These attributes will, however, be hidden since the network will form these relationships on its own. This is why the neural network sometimes is called a "black box".

A neural network consists of a collection of connected units called "neurons". Each neuron is connected through "synapses", which are used for signaling. A signal to the receiving neuron will be analyzed and sent forwarded to the neurons it can send to (Ng, Syllabus and Course Schedule, 2018b, p. 2).

For a neural network with 3 inputs, 3 hidden units, and one output layer would be illustrated as such:

Figure 3.1: A Simple Neural Network



Source: (Ng, Coursera, 2018a)

Where layer 1 is the input layer, layer 2 is the hidden layer, and layer 3 is the output layer. The arrows act as "weights" or the neural networks parameters. The intuition of the sigmoid hypothesis output $h_\Theta(x)$ is: the estimated probability that y=1 on input x_n. Adding all these intermediate layers in neural networks allows for a more elegant production of interesting and more complex non-linear hypotheses that reflect the mutual relationships between the inputs (Ng, Coursera, 2018a).

3.2.1.1 Backpropagation (Deep Learning)

The method described above is called forward propagation: the data is moved through the model and an output is received. The model's initial parameters, however, are arbitrary. Thus, "training" is needed, such that the model learns the optimal parameters in order to increase its prediction accuracy. This is done through backpropagation, where the data is moved the other way through the model.

To optimize the parameters, a "cost function" is used. A cost function takes the average difference between all the results of the hypothesis with inputs from x's and the actual output y's, or the difference between the predicted value and the actual value. The data is moved back and forth through the model as feedback in an iterative process in order to reduce the cost function by finetuning the parameters. If the cost function is minimized to 0, the model will perfectly fit the data. Thus, "Backpropagation" is neural-network terminology for minimizing the cost function. The cost function for a neural network is slightly complicated because one must account for the multiple output nodes. This is described in more detail in the appendix together with the neural network representation and notation.

The above model is a simple version of the neural network. In fact, neural networks can be "deeper" by having numerous units and layers, which is the intuition behind the term "deep learning". The RNTN used in this thesis is based on the same deep learning principles.

3.2.2 Word Vector Representation

The section describes word vector representation, which is a way of representing the semantics of words in a computer. This section is relevant because the RNTN makes use of the principles behind word vectors to build sentence vectors that express the meaning of sentences.

Representation of the meaning of words in a computer is a challenging task. When we want information or help from a person, we use words to make a request or describe a problem, and the person replies with words. Unfortunately, computers do not understand human language, so we are forced to use artificial languages and unnatural user interfaces (Turney & Pantel, 2010).

A part of the NLP area regard words as statistically independent (Socher & Manning, Natural Language Processing with Deep Learning, 2018). The problem with this is that it is difficult to accurately compute word similarity. In these terms, a word is a vector with one 1 and a lot of zeroes (equal to the size of the dictionary). For example:

$Motel$: $[0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad \dots \quad 0_n]$

$Hotel$: $[0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad \dots \quad 0_n]$

This is called a "one-hot" representation. The problem with this representation is, that there is no notion of similarity, because of the symbolic representation, even though motel and hotel are very similar in meaning. The words have no relationship with each other - each word is a notion to itself. Therefore, a way to encode meaning into the vectors would be beneficial. A way to overcome this is by using distributional similarity-based representations or semantic vector spaces. The essence of this idea is that statistical patterns of human word usage can be used to figure out what people mean (Turney & Pantel, 2010).

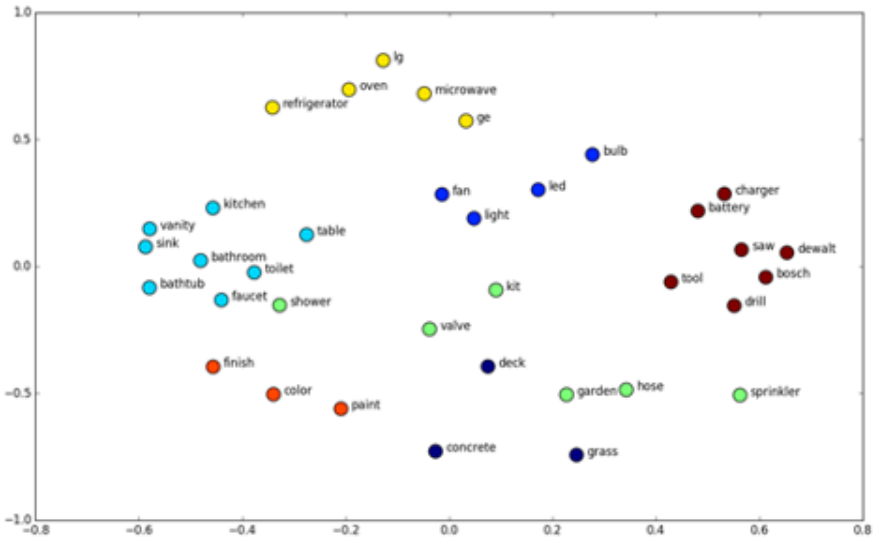
The idea of semantic vector spaces is to represent each word in a sentence as a point in space (a vector in a vector space). Points (word vectors) that are close together in this space are semantically similar and points (word vectors) that are far apart are semantically distant (Turney & Pantel, 2010). Semantics here means in a general sense as the meaning of a word.

The dominant approach in semantic vector spaces uses distributional similarities of single words. Often, co-occurrence statistics of a word and its context are used to describe each word, also called the latent relation hypothesis (Turney & Pantel, 2010) (Baroni & Lenci, 2010). Variations of this idea use more complex frequencies such as how often a word appears in a certain context (Lapata & Padó, 2007) (Erk & Padó, 2008). However, distributional vectors often do not properly capture the differences in antonyms since those often have similar contexts. One possibility to remedy this is to use neural word vectors (Bengio, Ducharme, Vincent, & Jauvin, 2003).

These vectors can be trained in an unsupervised fashion to capture distributional similarities (Collobert & Weston, 2008) (Huang, Socher, Manning, & Ng, 2012) but then also be fine-tuned and trained to specific tasks such as sentiment detection (Socher, Pennington, Huang, Ng, & Manning, 2011).

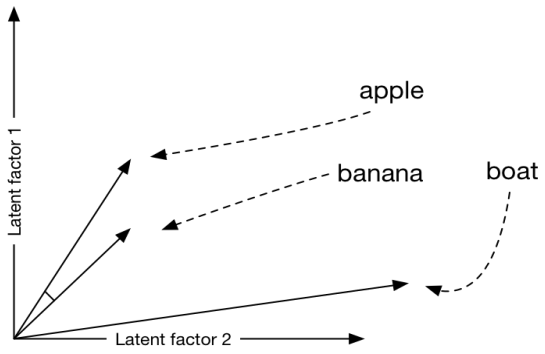
The following are illustrations of semantic word vectors:

Figure 3.2: Semantic Word Vectors In 2D Vector Space



Source: (Lynn, 2018)

Figure 3.3: Semantic Word Vectors In 2D Vector Space



Source: (Morrison, 2015)

It is seen that words with overall the same theme on cluster together. The reason for this is that words that appear in similar context often turn out to have a similar meaning. Thus, semantic word vectors are a way to obtain a form of similarity between words. The benefit of these learned word vectors is the ability to classify words and phrases

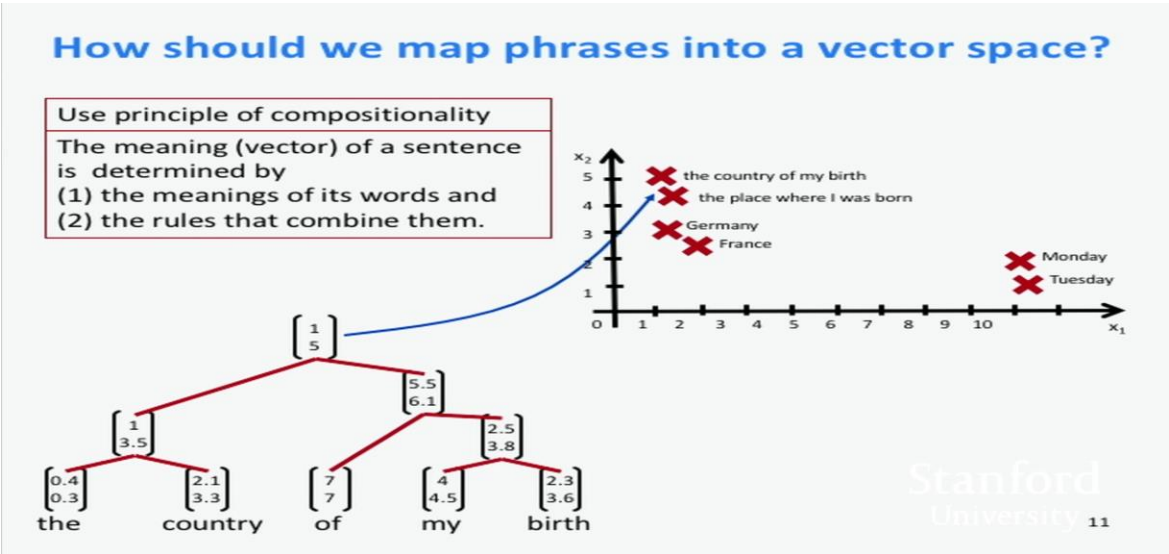
accurately. As seen in Figure 3.2, one example is that words for different tools cluster together. The implication of this is that classifying tool words should be possible since these words have similar vectors.

The RNTN model in this paper uses purely supervised word representations learned entirely on the new Stanford Sentiment Treebank (Socher, et al., 2013).

3.2.3 The Composition of Word Vectors into Sentence Vectors

In general, it seems that for a computer to understand sentences, there is a need to have models that have the capability for semantic compositionality. Compositionality means that you put together smaller pieces into larger pieces and work out the meaning of those larger pieces. The goal is to take bigger phrases and stick them into a vector space and represent their semantic similarity. For words, there is a big lexicon of words and the ability to learn a meaning representation for each one of them. That's not possible for phrases and sentences because there is an infinite amount of different of them in the English language, meaning it would be needed to calculate and store a vector for each phrase. An idea is to compose the meaning out of a phrase. A semantic composition can be achieved by combining the word vectors recursively. The goal is to obtain the meaning (vector) of a sentence is through the meanings of its words and the rules that combine them. For example, for the sentence: "the country of my birth" the goal is to combine the two words "my birth" have a meaning for that phrase, a meaning for the phrase "the country", and keep on calculating up and get a meaning for the whole phrase, which is then represented in the vector space. An illustration of this is shown in Figure 3.4:

Figure 3.4: Mapping Of Phrases into Vector Space



Source: (Socher & Manning, 2018)

The recursive neural network provides an architecture for jointly parsing natural language and learning vector space representations for variable-sized inputs. These networks can additionally induce distributed feature representations for unseen phrases and provide syntactic information to accurately predict phrase structure trees (Socher, Manning, & Ng, 2010). This means that the model can also be used to predict the structure of sentences it hasn't seen before.

Socher et al. (2011) use this structure for accurately parsing natural language and Socher et al. (2012) use the structure as a matrix-vector RNN. The main idea of the MV-RNN is to represent every word and longer phrase in a parse tree as both a vector and a matrix. When two constituents are combined the matrix of one of one constituent is multiplied by the vector of the other and vice versa. Hence, the compositional function is different according to the words that participate in it.

Socher et al (Socher, et al., 2013) explain that one problem with the MV-RNN is that the number of parameters ones becomes very large since there are many specific inputs (one for each word in the vocabulary). They posit that it would be more plausible if there was a single powerful composition function with a fixed number of parameters that can aggregate meaning from smaller elements. The standard RNN (Socher, Lin, Ng, & Manning, 2011) would be a good candidate for such a function. However, in the standard RNN, the input vectors only implicitly interact through their combined phrase vector. A more direct, multiplicative interaction would allow the model to have greater relations between the input vectors. Thus, they propose a new model called the Recursive Neural Tensor Network (RNTN), which has a powerful tensor-based composition function. The main idea is to use the same, tensor-based composition function for all nodes, which allows for interaction between the input vectors at the bottom level.

The strength of the RNTN model's method lies in its ability to identify a specific type of phrase composition and relate it to other types of phrase compositions. In other words, the tensor component is able to combine input vectors such that there is a definition of them in vector space and directly relate the vector to similar vectors. In that way, it builds the sentence structure and sentiment classification from the relations.
The combination of the RNTN model and the Stanford Sentiment Treebank results in a system for single sentence sentiment detection that pushes state of the art for positive/negative sentence classification. Besides that, it captures negation of different sentiments and scope more accurately than previous models (Socher, et al., 2013). These characteristics make it an ideal model for incorporating grammar into sentiment analysis of narratives in 10-K reports.

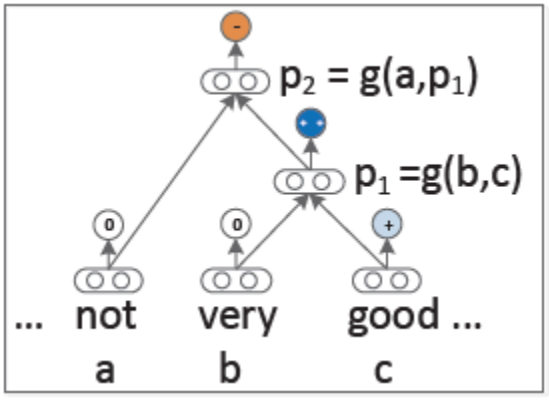
3.2.4 The Stanford Sentiment Treebank

The RNTN utilizes a large labelled compositionality resource called the "Stanford Sentiment Treebank", which includes labels for every syntactically plausible phrase in thousands of sentences. Specifically, it consists of 11.855 single sentences extracted from movie reviews. The sentences were parsed with the Stanford parser (Klein & Manning, 2003) and includes a total of 215.154 unique phrases from those parse trees, each annotated by 3 human judges. The parse tree structure allows for complete analysis of the compositional effects of sentiment in language through the RNTN model that is trained to accurately predict the compositional semantic effects present in the corpus.

3.2.5 A Mathematical Explanation of the Recursive Neural Tensor Network (RNTN)

Each word is represented as a d-dimensional vector. All the word vectors are stacked in the word embedding matrix $L \in R^{d*|V|}$, where V is the size of the vocabulary. Initially, the word vectors will be random but the L matrix is seen as a parameter that is trained jointly with the compositionality models.

Figure 3.5: Trigram of Phrase Construction

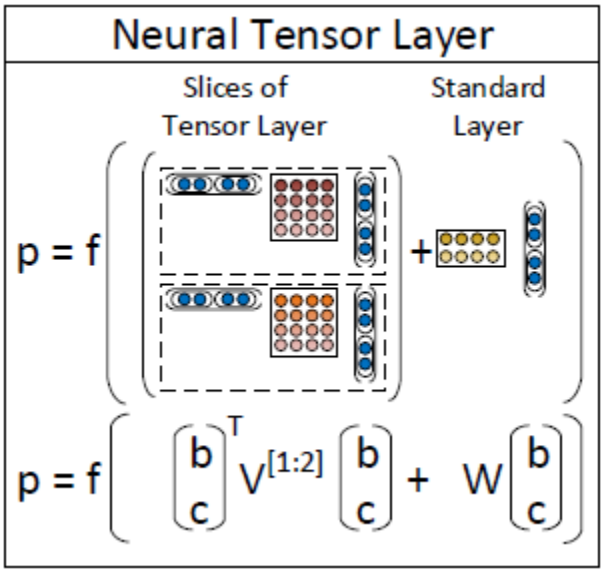


Source: (Socher, et al., 2013, p. 4)

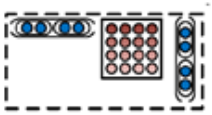
The word vectors are used as parameters to optimize and as feature inputs to a logistic classifier. The logistic function will classify each word into one of five different sentiment classes: very negative, negative, neutral, positive, and very positive. It will do this, by computing the probability of the label given the word vector via: $y^a = softmax(W_s a)$,

where $W_s \in R^{5*d}$ is the classification matrix, which has the dimension: 5 sentiment classes * word vector dimension, and a is the given word vector. For the example in the Figure 3.5, this process is repeated for word vectors b and c. The main task of the model is to compute the hidden parent vectors in a $p_i \in R^d$ in a bottom up fashion as illustrated in Figure 3.5.

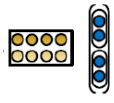
Figure 3.6: A Single Tensor Layer



Source: (Socher, et al., 2013, p. 7)

Figure 3.6 shows a single tensor layer. The concatenation of the word vectors b and c gives the bilinear form: $\begin{bmatrix} b \\ c \end{bmatrix}^T V^{[1:2]} \begin{bmatrix} b \\ c \end{bmatrix} =$  , where the output is a single number. The parent representation must have the same number of dimensions as each of the children vectors. In this example, the children are 2-dimensional, thus, the output must also be 2 dimensional. Generally, if the children are n-dimensional, the output should be n-dimensional also. Therefore, another tensor layer "slice" is added - each slice will then output one number for one of the hidden units of the resulting parent vector. The following vectorized notation shows the output of a tensor product defined as $h \in R^d$, and the output for each slice of tensor layer defined as $h_i \in R^d$:

$h = \left(\begin{bmatrix} b \\ c \end{bmatrix}^T V^{[1:2]} \begin{bmatrix} b \\ c \end{bmatrix}\right); h_i = \left(\begin{bmatrix} b \\ c \end{bmatrix}^T V^{[i]} \begin{bmatrix} b \\ c \end{bmatrix}\right)$, where $V^{[1:d]} \in R^{2d*2d*d}$ and $V^{[i]} \in R^{d*d}$ are multiple bilinear forms that transform the word and phrase vectors into vectors that reflect the compositional value of the two.

Then the standard recursive neural network is added: $W \begin{bmatrix} b \\ c \end{bmatrix} =$  , followed by the element-wise non-linearity (softmax) function, and p will be: $p = f\left(\begin{bmatrix} b \\ c \end{bmatrix}^T V^{[1:2]} \begin{bmatrix} b \\ c \end{bmatrix} + W \begin{bmatrix} b \\ c \end{bmatrix}\right)$.,

This is now a powerful model that allows for different interactions. Each of the $V^{[1:2]}\begin{bmatrix} b \\ c \end{bmatrix}$ parameters are equal to one hyper frame in the vector space and capture one way of how each of these word vectors interact. On top of this parent vector, a logistic regression is used to train on all the parameters in a backpropagation style.

3.2.5.1 RNTN Backpropagation

Training the RNTN model is similar to training a simple neural network through backpropagation, however, it is slightly more complicated given its tree structure. In general, each node has a softmax classifier that is trained through the vector representation to predict a target vector t. The target vector t is assumed to be a one-hot vector with a number of classes (sentiments) as length (C) - a 1 at the correct label and all other values to be 0. The predicted vector y will be the probabilities of the classes and is a product of the classifiers.

The goal is to maximize the probability of the correct sentiment prediction. This is done by minimizing the cross-entropy error between the predicted distribution $y^i \in R^{C*1}$ at node i and the target vector $t^i \in R^{C*1}$ and that node (Socher, et al., 2013). A full explanation and mathematical notation of the backpropagation in the RNTN model is included in Appendix 4.

Chapter 4 – Extraction of Sentiment

This chapter will explain how the 10-K narratives will be transformed into a quantitative sentiment score by the programming behind the BoW model and the software the RNTN uses. In addition, an explanation is given of how these classifications have been transformed into independent variables of concern. Lastly, a discussion of the pitfalls there might be when interpreting the BoW's and RNTN's sentiment scores in the further analysis is presented.

4.1 The Bag of Words Output

The following sections describe how the BoW model processes the 10-K reports into sentiment scores. First, the choice of dictionary is discussed, followed by the process of the programming that transforms the 10-K reports into sentiment scores.

4.1.1 Choice of Dictionary

The choice of dictionary for the BoW approach is important because, as mentioned in Section 3.1, it has a large say in what content will be extracted and the meaning that is inferred. The dictionary is a reference point that is used to decide how to assign words to different categories such as a positive or negative sentiment. The programming used for the BoW approach will be described in the next section, while in this section the choice of dictionary for the sentiment classification of the 10-K reports will be discussed.

The use of a dictionary has three advantages:

1. It makes sure that research subjectivity is avoided when classifying the sentiment of the different words.
2. The method is easy to scale, since it is possible through programming, to make frequency counts of a large scale.
3. The dictionaries are available to the public, which allows other researchers to replicate the study.

(Loughran & McDonald, 2016, p. 1200)

The most used dictionaries in accounting and finance relate to four different dictionaries, which are:

1. The Henry Word List
2. Harvard GI Word List
3. Diction Optimism and Pessimism Word List
4. Loughran and McDonald Word List

(Loughran & McDonald, 2016, p. 1200) (Henry & Leone, 2014, p. 37)

When choosing between these word lists it is important to evaluate what the different word lists can provide compared to the intentional use of the word list. The different word list will therefore mainly be evaluated based on two criterions which are:

1. Are they created with a business context in mind?
2. Is the amount of words in the word list comprehensive?

The first criterion is chosen since the sentiment classification of words will vary depending on the context they are used. In this study, which analyses 10-K reports, using a word list that is specific to a business context will diminish the classification errors giving more consistent results. The reason for the second criterion is that if the word list is not comprehensive enough, then a lot of frequently used words in accounting will be missing and the model might not classify the sentiment of the 10-K report correctly.

Based on these criteria the different strengths and weaknesses of the word lists will be discussed and based on this, the dictionary that is used in the BoW model will be chosen.

The four word lists will be evaluated based on the abovementioned criteria, and the most optimal for the sentiment analysis with the BoW method will be chosen.

The Henry Word List

The Henry word list is one of the first word lists created specifically for financial texts. It was created by examining earnings press releases in the telecommunication and computer-service industries.
The word list contains 105 positive and 85 negative words and the different synonyms to these words (Henry, 2008, pp. 387-388).
The focus of the word list is only on corporate earnings press releases, and it is in that regard restricted to a limited research area. The effect is it only contains a limited number of words, and a lot of frequently used words in accounting are missing from the word list. Examples of this could be loss, losses, and impairment (Henry, 2008, pp. 387-388) (Loughran & McDonald, 2016, p. 1201). This is an obvious weakness of the word list since it can give wrong results, given that if a lot of frequently used words in accounting are missing, the words will not be included in the frequency count, and the sentiment of the text can be misstated. There is furthermore some ambiguity in some of the words included in the Henry word list. A word such as "increased" is labelled as positive and "decreased" is labelled as negative. These two words are not necessarily positive or negative, but it depends on which context they are used. If revenue is taken as an example, it is positive if it increases, but if cost increases it could be a negative thing (Henry, 2008, p. 388).
It can, therefore, be argued that the Henry Word List word list has some problems that would make it unreliable to

use. The positive thing about the word list, on the other hand, is that it is created with a business context in mind even though it is only based on earnings press releases.

Harvard GI Word List

Together with the Diction Optimism and Pessimism Word List, the Harvard GI Word List were some of the first dictionaries available. These dictionaries were therefore used in most of the initial research on textual analysis in finance and accounting. The GI Word List consists of 4.187 negative words (Loughran & McDonald, 2016, p. 1201), however, the dictionary was not created with an accounting and finance context in mind, and it is therefore reported by Loughran and McDonald that 75% of the negative words do not have a negative meaning in a business context. Examples of this could, for example, be that tax, cost, capital, board, and liability are classified as negative. These words are used extensively in financial reporting, and usually not in a negative way. This would increase the measurement error if this dictionary was used. Furthermore, Loughran and McDonald document that words such as crude, cancer or mine that are classified as negative will punish some industries, since oil, pharmaceutical, and mine companies might use these words extensively without having any negative meaning in this context. The Harvard GI Word List will, therefore, be prone to errors when used in a business context (Loughran & McDonald, 2016, p. 1203).

Diction Optimism and Pessimism Word List

The Diction Optimism and Pessimism Word List consists of 31 different dictionaries, which vary in size from 10 to 735 words. These dictionaries have been combined into 5 different lexical features called master variables, which are (Diction, 2015, p. 1):

1. Certainty
2. Activity
3. Optimism
4. Realism
5. Commonality

The variables have different definitions and it is especially the variable optimism that is interesting since this variable will explain to what degree a text is positive.
The variable is created by combining 6 dictionaries, which are praise, satisfaction, inspiration, blame, hardship, and denial, where the first three are positive and the last three are negative. When using this approach 686 optimism and 920 pessimism words are used (Diction, 2015, pp. 4-5).
When analyzing the different words included in the separate dictionaries it can be argued that some of the words included are not accurate when used in a business context, which makes sense since, like the Harvard GI Word List,

it was not created with a business context in mind (Henry & Leone, 2014, p. 37). To mention some examples, it can be argued that words such as respect, power, and trust are not truly positive. At the same time, the included negative words: no, not, without and gross are not truly negative in a business context (Loughran & McDonald, 2015, pp. 1-2) (Diction, 2015, p. 7).

Loughran and McDonald Word List

The Loughran and McDonald Word List is a comprehensive word list that consists of 85.221 words. It was initially created in 2009 and has been regularly updated with the last update made in 2017 (McDonald, 2018b). The word list is based on the 4.0 edition of the 2of12inf dictionary, which includes around 82.000 American English words. The 2of12inf word list contains 12 different dictionaries, which includes inflections, but not abbreviations, acronyms or names (Beale, 2018).

To create the final Loughran and McDonald Word List all the different versions of 10-K and 10-Q filings from 1994 until 2016 are examined. The words that are not included in the 2of12inf word list and that occur more than 50 times in these filings will then be added to the Loughran and McDonald Word List.

In the end, 6 different sentiment categories are defined, which are:

1. Negative
2. Positive
3. Uncertainty
4. Litigious
5. Modal
6. Constraining

These different sentiment categories have been defined based on their most likely interpretation when used in a business context (Loughran & McDonald, 2016, p. 1204). When looking at the positive and negative words the list is rather comprehensive since it includes 2355 negative and 354 positive words and should, therefore, be able to extract the sentiment from the 10-K filings. Even though it is a rather comprehensive list of positive and negative words, it has been argued that it is a weakness that the word list has not defined words such as "up" and "increase" as positive. On the other hand, it can be argued that these words are ambiguous and could have a positive or negative meaning depending on the context (Henry & Leone, 2014, p. 38).

Given the characteristics of the Loughran and McDonald word list, it can be assumed that it is rather comprehensive compared to the other word lists. It is furthermore produced by using 10-K and 10-Q reports and with a business context in mind. The Loughran and McDonald Word List is furthermore updated on a regular basis, which makes sure that it is kept up to date when the language is changing.

Reasons for Choosing the Loughran and McDonald Word List

In the sections above the four most used dictionaries in accounting and finance have been discussed based on the two evaluation criteria.
Based on this discussion and criteria the Loughran and McDonald Word List has been chosen.
The first examined criterion was whether the word list was created with a business context in mind. Out of the four-word lists, it is only the Loughran and McDonald Word List and the Henry Word List that are created with a business context in mind. However, while the Henry Word List is based on earnings press releases, the Loughran and McDonald Word List is based on 10-K and 10-Q filings, which gives it an edge since it arguably better covers the spectrum of the words in the narratives that are used as data in this thesis.
The second criterion was, whether the words in the word lists were comprehensive enough. The Henry Word List consists of 105 positive and 85 negative words. The word list is therefore rather small compared to the Loughran and McDonald Word List which consists of 2355 negative and 354 positive words. The effect of this is that compared to the Loughran and McDonald Word List the Henry Word List is not as effective at identifying and classifying words and therefore might give weaker results.
Based on the above considerations, the most convincing dictionary in the context of analyzing the sentiment of 10-K reports is determined to be the Loughran and McDonald Word List.

4.1.2 The Bag of Words (BoW) Model's Sentiment Scores

This section explains how the BoW model transforms text into sentiment scores using the programming language Python and the Loughran and McDonald Word List.
The 10-K reports have initially been preprocessed as described in Section 5.1.2.2. Before these preprocessed 10-K reports can be used as input into the BoW model, however, the text of these reports has to be converted into a vector of words. These vectors are created by counting the frequency of all the distinct words that are represented in the 10-K report. Each word in the model will, therefore, be represented by their occurrence in the text (Sarkar, 2016, pp. 178-179).
To create these vectors of words the 10-K report must be tokenized, which is a preprocessing approach, where the different characters of the text, such as letters, are identified. The identification of words is done by finding a collection of characters that occur between word boundaries in the text, where word boundaries in this context are defined as blank spaces. The reason for tokenization is that the characters in an electronic text are a linear sequence of characters, words, or phrases, and to get meaningful input into the model, the text must be converted into tokens, which roughly corresponds to words (Trim, 2013).

In order for these tokens to be applied as input into the BoW model, the tokens have been through additional processing. The processing of the tokens consists of four steps that will be described next.

First, all letters in the text have been changed to capital letters since lower- or upper-case letters are not informative in this analysis of sentiment. This will also make sure that the words in the word list are recognized since these are case sensitive.
The second thing that has been done is to remove all stop words. Stop words are words that appear often and provide no discriminatory power in information retrieval (Henry, 2008, p. 400). The excluded stop words are shown in Appendix 3.
The third adjustment that has been made is that hyphenated words are split, which means that the model does not take hyphenated words into account. To circumvent this the Loughran and McDonald Word List includes the hyphenated words both separately and as one word (Loughran & McDonald, 2016, p. 1216).
Finally, all numbers are removed.
After the 10-K report has been through these different steps of processing, the tokens are collected into a vector of words. This vector of words will be compared with the with the Loughran and McDonald Word List.
The Loughran and McDonald word list is first of all used to determine whether the words included in the model are actual words, since tokens that are present in the 10-K report will only be included if these tokens are also present in the Loughran and McDonald Word List. In addition, and arguably more importantly, it is used to assign words into different categories such as if the sentiment of a word is positive or negative, and by doing that extract the sentiment from the 10-K reports. The output of the BoW will, therefore, be a count of the number of words that are included in the report, and a count of how many of these words that are defined as positive or negative based on the Loughran and McDonald Word List.

Finally, a sentiment score of the positive and negative words will be calculated. This sentiment is calculated as the percentage of positively or negatively classified words out of the total number of words in the 10-K report.

4.2 Stanford CoreNLP Output

This section will explain how the Stanford CoreNLP software transforms the 10-K reports into sentiment scores and how the scores are accumulated on a document level.

Stanford CoreNLP is the software from which the RNTN is executed from. It is a modern, regularly updated package, with the overall highest quality text analytics. Through an integrated NLP toolkit with a broad range of grammatical analysis tools, its goal is to make it very easy to apply a bunch of linguistic analysis tools to a piece of text. Its many applications include the following:

- Giving the base forms of words
- Parts of speech tagging
- Indicate ting which noun phrases that refer to the same entities

- Determining whether the words are names of, for example, companies or people
- Normalization of dates, time and numeric quantities
- Marking up the structure of sentences in terms of phrases and syntactic dependencies
- The ability to extract particular or open-class relations between entity mentions
- Last and most importantly, it is able to indicate the sentiment of sentences

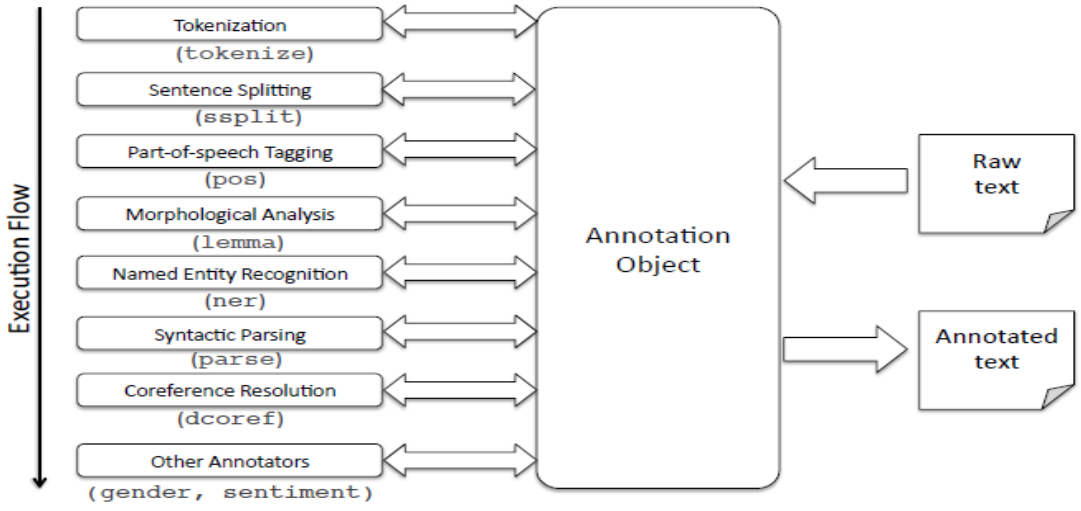
(Stanford, 2018a).

The Stanford CoreNLP software is a Java annotation pipeline framework. Some of the motivations for making this architecture was to provide support for:

- Quickly and painlessly getting linguistic annotations for text.
- To have a minimal conceptual footprint, so the system is easy to learn.
- To provide a lightweight framework.

One of the key design goals was to make it very simple to set up across multiple platforms and run processing pipelines, from either an API (Application Programming Interface) or the command-line, with minimal configuration code. Keeping it simple, provides new users with a very good initial experience (Manning, et al., 2014). The additional benefit in this study is that it allows for easy replication in future studies, something that Loughran and McDonald advocate for (Loughran & McDonald, 2016, p. 1189). Figure 4.1 illustrates the overall system architecture:

Figure 4.1: Stanford CoreNLP Pipeline Framework



Source: (Manning, et al., 2014)

Raw text is put into an Annotation Object and then a sequence of Annotators adds information in an analysis pipeline. The output can be in XML or plain text forms (Manning, et al., 2014).

To get the sentiment of the sentences from the 10-K narratives, the sentiment annotator, which implements Socher et al's (2013) sentiment model, is used. The sentiment annotator, however, is dependent on the tokenize, sentence split, part of speech, lemma and parse annotators. The following is a description of these annotators and their process of transforming the 10-K reports into sentiment scores.

By locating word boundaries, the tokenizer divides the text into a sequence of tokens, much like the process in the BoW model explained in Section 4.1.2. Sentence splitting is a deterministic consequence of tokenization. The Stanford Core NLP software determines that a sentence ends when a sentence-ending character (., !, or ?) is found, which is not grouped with other characters into a token (such as for an abbreviation or number), though it may still include a few tokens that can follow a sentence ending character as part of the same sentence (such as quotes and brackets) (Stanford, 2018c).

A Part-Of-Speech Tagger (POS Tagger) is a piece of software that reads text in some language and assigns parts of speech to each word, such as noun, verb, adjective, and so on. The software used in the Stanford CoreNLP is a Java implementation of the log-linear part-of-speech taggers described in Toutanova et al (2003) (Stanford, 2018b). Lemmatization aims to remove inflectional endings and to return the base or dictionary form of a word, which is known as the lemma. For example, if confronted with the token *saw*, lemmatization would attempt to return either *see* or *saw* depending on whether the use of the token was a verb or a noun (Stanford, 2009).

A natural language parser is a program that works out the grammatical structure of sentences, for instance, which groups of words go together (as "phrases") and which words are the subject or object of a verb. Probabilistic parsers such as the Stanford Parser use knowledge of language gained from hand-parsed sentences to try to produce the most likely analysis of new sentences. These types of statistical parsers still make some mistakes, but commonly work rather well. Their development was one of the biggest breakthroughs in natural language processing in the 1990s. The original version of the Stanford Parser was mainly written by Dan Klein, with support code and linguistic grammar development by Christopher Manning (2003) (Stanford, 2018e).

Finally, based on the above input, the sentiment annotator will, by implementing the RNTN, classify the sentiment of the sentences. For each text file containing a 10-K report, the Stanford CoreNLP software will run the text through the annotators and in the end provide an output text file, which will contain a sentiment classification for each sentence in the 10-K report. Through programming in Python, the sum of the different sentence classifications in each output file are extracted to Microsoft's Excel. In doing so, there is a count for every negatively and positively classified sentence for each 10-K report, which is used as input for further analysis.

4.3 Limitations of the Sentiment Scores

This section will describe the different pitfalls there might be when interpreting the BoW's and RNTN's sentiment scores of the 10-K reports.

4.3.1 Limitations of the RNTN's Sentiment Score

Although the Stanford Sentiment Treebank is based upon a vast number of sentences and phrases, it poses some limitations for its usage in financial texts because of the differences in the language used in movie reviews and financial documents. The main difference is that most people leaving movie reviews either hate or love the movie and will tend to use colourful language in that regard, rather than leaving a neutral review analyzing the technical terms such as the financing of the movie. This is a problem because for the RNTN to accurately predict compositional semantic effects, it needs to have seen the sentence or a sentence similar to it to be able to classify its sentiment.

However, given the level of detail and size of the corpus, it is estimated, that despite the language domain mismatch, the RNTN will still be able to a large degree suggest distributed feature representations for the unseen phrases in the 10-K and provide syntactic information to predict its phrase structure trees and sentiment.

An initial test was made to evaluate the performance of the RNTN model. A sample of text from Danske Bank's and Tesla's annual report was analyzed. Out of 107 sentences, the models classified 22 as positive, 83 as negative and 2 as neutral. These results indicate a high negative bias. Upon some research on the internet, it appears that the model has a tendency of classifying sentences with the characteristic of being somewhat technical and otherwise neutral as negative. This poses a problem since the model's negative score becomes dubious. It is unsure whether a high negative sentence ratio in a 10-K report is due to overly negative language use, or because of a large amount of neutral technical terminology. Given that 10-K reports are characterized by a large amount of neutral technical terminology, the use of the RNTN model seems discouraging. However, one uplifting aspect of this model is that the sentences it did identify as positive in the test were to a large degree positive (this assessment was made through face validity). Out of the 22 positive sentences, 18 of them were deemed to be true positive, which corresponds to an accuracy of roughly 80%. One example of a sentence it deemed positive is: "At the same time, we invested further in the development of our business and launched a number of new and innovative products and solutions aimed at making both daily banking and important financial decisions easier for our customers" (Danske Bank Group, 2018, p. 4).

Since the model can't meaningfully interpret much of the neutral technical terminology it limits its use to an analysis of parts of the 10-K report where an opinion is clearly stated through subjective language, such as the sentence above. Given, however, the fact that analysis of stock returns and sentiment will be based on the change from the

previous year this might benefit the predictive power of the RNTN, which is described in Section 5.2.3. The reason for this is that the technical language would likely be the same across years for the same firm, meaning that the negativity ratio that is explained by technical language is recurring.

4.3.2 Limitations of the BoW Model's Sentiment Score

When analyzing sentiment with the BoWs approach, the three main measures are the percentage of negative or positive words and the net sentiment score, which is the percentage of positive and negative words subtracted from each other. These measures have all been used in prior research, however, not without some challenges. Especially the positive sentiment score can pose a problem when included in the analysis. The reason for this is that companies have a tendency to frame a negative statement, by negating a sentence with many positive words. An example of this can be: "in 2007, the global automotive industry continued to show strong sales and revenue growth, however, we experienced an overall reduction in our net income". In that way, the positive words "strong" and "growth" can be used to frame a negative statement. The opposite case is rarely true since a company rarely negates a negative statement to make it positive. Negative sentiment scores will, therefore, be less ambiguous, since they are not influenced by framing issues (Loughran & McDonald, 2016, p. 1217).

In the end, careful consideration must be made when interpreting the results of the models. Both models are flawed in the sense that even though the BoW model can make use of a financial dictionary, it ignores the sequence of words, which is not sensible from a linguistic point of view. This is especially evident with negations. On the other hand, the RNTN is trained on language that is a different domain than finance, which does not make it as practical either as it could be. The main pitfall is that there is a large negativity bias since the model categorizes neutral technical sentences as negative. This bias can be somewhat rectified by examining the difference of the sentiment scores for the same firm between years.

Chapter 5 – Data and Regressions

Chapter 3 and 4 described the theory behind the BoW and RNTN models, how the 10-K narratives will be transformed into a quantitative sentiment score by the software the BoW and RNTN uses, and how these classifications have been transformed into independent variables of concern. The aim of this chapter is to discuss and explain the choices of data of the process of the analysis behind the results in the thesis. The first section will describe the different data sources that have been used, the choices of what data to base the analysis upon and how this data has been retrieved. The second section will discuss the challenges of parsing the 10-K reports. Finally, the analysis is in focus, where the methodological choices of the regression between the sentiment scores and the stock returns will be presented.

5.1 Data

The retrieval of data and making it usable for analysis purposes has been one of the tasks that have taken the most time and effort in this thesis. An extensive amount of time has been used to make sure that data is of the right quality and with a minimal amount of bias. This section will be structured in the following way:
The first section will give a description of the different data sources that have been used. The second section will evaluate the data. This section will be split into a section that describes the different problems which have occurred when using the 10-K reports, and what choices that have been made to overcome these. The other section will describe how the S&P 500 index is composed and how the stock data is used and collected.

5.1.1 Data Sources

The data sources concern the databases that have been used to retrieve 10-K reports and various market-related data.

WRDS

WRDS is a research platform that gives access to databases across multiple disciplines such as accounting, economics, and finance. This platform was used to retrieve market information for the different companies in the S&P 500 index. To retrieve the needed market information two different databases were used.
First of all, the Compustat database was used. This database was used to retrieve the constituent list of which companies were included in the S&P 500 index. The constituent list describes which companies have entered and been deleted from the index in a period of time.
After the constituents list had been retrieved, the returns of the companies included in the S&P 500 index had to be found. This was done in the CRSP database. The reason for using the CRSP database is that it is a provider of historical stock market data and is frequently used in scholarly research. It is therefore regarded as a trusted source (CRSP, 2018a). It was also possible to get stock data from Compustat but to calculate returns the stock price had to be adjusted for splits and dividends. Compustat has an adjustment factor to do this, but it was error proven, hence

the returns could not be trusted. It was therefore chosen that the stock information to calculate returns should be retrieved from CRSP.
The problem of using both Compustat and CRSP data is that the two databases do not use the same identification codes for the different companies. It was, therefore, necessary to find another way to merge these different datasets. To circumvent the problem of different identification codes, CRSP has made a database available, where the different identification codes from Compustat and CRSP can be merged, which was therefore used. How this merging process was done is described in Section 5.2.4.
Finally, accounting data had to be retrieved for all companies included in the S&P 500. The accounting data was retrieved from Compustat, and since it was a database that had already been used, there were no problems, when the accounting data had to be merged to the other datasets.

EDGAR

When companies file their 10-K reports to SEC, they are gathered in the EDGAR database and made available to investors to download (Loughran & McDonald, 2017, p. 1).
This download process of 10-K reports has been done by Professor Bill McDonald, who has downloaded all these filings and made them available at a cloud service (McDonald, 2018a).
The available files in this cloud service have been downloaded and are used as input into the Stanford CoreNLP and the BoW model.

5.1.2 10-K Reports and their Applicability for Automated Textual Analysis

This section gives a description of the different considerations that are necessary when using the 10-K report. First of all, it is important to be aware of what the differences between the different 10-K reports are, and whether they should be included. Secondly, it is important to be aware of how the narratives are parsed from the 10-K reports and what challenges there might arise in doing so since the process of doing this will have a major influence on the results.

5.1.2.1 Types of 10-K Reports

The 10-K is, as described in Section 2.4, the standardized form that the annual reports are filed in, however, there are other types of filings that relate to the 10-K filing that are important to know. The two other filings are the 10-K/A and 10-KT (McDonald, 2018c). The 10-K/A form is a report which is used if there are any amendments to the original 10-K report. This form could be used if there is any omitted material information or if the company chooses to file the financial schedules as an amendment (EY, 2017, p. 27). The last report type is a 10-KT report which is used if the company changes its fiscal year-end. It will be used instead of or in addition to the original 10-K report and afterward, the company will file a regular 10-K report. The 10-KT will, therefore, be used if there is no 10-K

filing in the current year.
The reports that will be used are therefore the 10-K reports, but if a 10-K report has not been filed in the current year, then the 10-KT file will be used if this has been filed instead.

5.1.2.2 Parsing of 10-K Reports

The 10-K reports have been through two types of parsing. It was described in Section 5.1.1 that the 10-K reports have been downloaded from a cloud service, which was made available by Professor Bill McDonald. The reports downloaded from this cloud service has been through an initial parsing, before a second parsing of these reports was performed.
This parsing was done to make the 10-K reports usable for textual analysis, and to extract the narrative of the 10-K reports.

Initial Parsing of 10-K Filings

This section will describe the initial parsing of the 10-K reports by Bill McDonald. The overall purpose of this parsing is to make the reports usable for textual analysis.
The largest filings that are downloaded from the EDGAR database exceed 400 Megabyte(MB). It is, therefore, necessary to decrease this file size to make them usable as input into the different models. When an EDGAR text filing such as the 10-K report is used, it consists of more than just the plain text. It also consists of embedded markup tags such as HTML, XBRL or XML code, tables, and ASCII-encoded graphics. Since the focus is on the textual content of the document these items have been excluded from the final text. On top of the items mentioned above, everything from the beginning of the original document through the markup tag </SEC-HEADER> is removed, which means that different information such as the business address and mail address of the company is removed (McDonald, 2018c).
Finally, a header tag and an exhibit tag are added to the document. By doing this the different exhibits and the information regarding the file is easily removed afterward (McDonald, Stage One 10-X Parse Data, 2018c).

Second Parsing of 10-K Reports

After the first parsing was made by Bill McDonald, the data was parsed further into what is the final text that is used as input into the RNTN and the BoW model. This was done by examining 50 different reports to evaluate if the data quality could be improved and remove noise from the text. By doing this the following was identified and excluded as well:

- Most reports have page numberings on a single line. These numbers and the single line were therefore removed. This was done for both Roman numerals and Arabic numerals.
- Some page numbers had markers around the page numbers, which was therefore removed.
- All tabulating characters were removed, which are often used in the tables of contents.

- All appearances of "PART" in upper case was removed since companies use "PART" to begin a new part of the report.

On top of these specific characters, that were removed, the two markup tags used by Bill McDonald representing the header and exhibits were removed. The final text does therefore not contain any exhibits (Appendix 6).

5.1.2.3 Challenges Using 10-K Reports

When 10-K reports are used for automated textual analysis, different challenges were encountered. These challenges and how they were treated will be described in this section.

Management Discussion & Analysis (MD&A) in 10-K Reports

The initial idea behind this thesis was to analyze the MD&A. The reason being that the MD&A is the section in the 10-K, where management is required to comment on their financial condition, changes in financial conditions and results of operations, and by doing this, they express their opinion about the future as well (EY, 2017, pp. 71-72). The purpose of the MD&A is therefore:

> *"to provide investors and other users information relevant to an assessment of the financial condition and results of operations of the registrant as determined by evaluating the amounts and certainty of cash flows from operations and from outside sources."*

Source: (EY, 2017, p. 72)

SEC furthermore gives a guidance of how to interpret the disclosure requirements of the MD&A in FR-72. This guidance especially states three objectives of the MD&A, which are:

1. A narrative explanation of the company's financial statements that enables investors to see the company through the eyes of management is required.
2. An enhancement of the overall financial disclosure and provide the context for analysis of financial information.
3. At last the company should provide information regarding the quality of, and potential variability of, the company's earnings and cash flow, which should make the investor able to ascertain the likelihood that past performance is indicative of future performance.

(U.S. Securities and Exchange Commission, 2003)

Based on these objectives, the MD&A should be a discussion and analysis of a company's business seen through the eyes of management. The reason is that management has a unique perspective on the business, which only they are able to present. In addition, the MD&A should not be a recitation of the financial statement in a narrative form or an uninformative series of technical responses to MD&A requirements, neither of which provides the important management perspective (U.S. Securities and Exchange Commission, 2003).

The MD&A is, therefore, the place to look in the 10-K when sentiment should be analyzed since management in this section will give their perspective on the most important matters to the company in narrative form.

Selection of Full 10-K Report

The optimal solution when analyzing sentiment in 10-K reports would, as described above, be to use the MD&A section of the 10-K. When the process of extracting the MD&A from the 10-K report was initialized, it was however made clear that the quality of data was too low when using this approach.

The data quality was low since there must be some logic that is consistent when a text is parsed. This logic has been made available for other items in the 10-K report with the use of XBRL, which is a reporting language that provides a computer-readable tag for each individual item of data. This XBRL tag is however not required for the MD&A section (EY, 2017, pp. 10-11). Because of this, there is no logic that is either consistent or present when the MD&A is to be parsed. Even though this was the case different parsing approaches were tried, but in the end, the result of the missing logic was that other items than the MD&A were parsed, and there was, therefore, no assurance that the analysis would be based purely on the MD&A. The data would therefore not be consistent or useful as input into the RNTN and the BoW model.

Bill McDonald, one of the leading researchers in the field of textual analysis in finance, was furthermore asked if he had parsed the MD&A of 10-K reports successfully. His answer was:

> *"Parsing MDA's accurately is, in my opinion, virtually impossible and yet everybody claims to do it. If firms all followed the standard rules in terms of the form structure it would not be difficult, but they do not. In addition, many times the MD&A is put into an exhibit which is introduced in section 7, sometimes with enough introductory comments to make it difficult to determine where the MD&A is. I am very skeptical of research that leans heavily on this parse, but I know it is frequently done."* (Appendix 1)

With regard to our own results of parsing the MD&A and Bill McDonald's opinion regarding the possibility of parsing the MD&A, it was chosen to make the analysis based on all items of the 10-K report. By using all items, the results will be consistent since the input will be reliable. However, by analyzing the whole 10-K report there will undeniably be more noise in the analysis since the sentiment scores will not only reflect favorable or unfavorable information about performance.

Inclusion of Items by Reference

There are different rules specifying how to include the different items in the 10-K report. The items in part 1 and part 2 can be included by reference, by referring to the annual report to shareholders (see Table 2.1). The requirement is that the incorporated part from the annual report to shareholders contains the same information that is required in the 10-K report.

It is also possible to incorporate part 3 by reference, where the companies refer to a proxy statement, which is information required by SEC that will make shareholders able to make informed decisions on matters that will be

brought up at annual stockholder meetings (U.S. Securities and Exchange Commission, 2009a, pp. 2-3).
This possibility for companies to refer to either the annual report to shareholders or to a proxy statement, might influence the results. It is therefore important to test, to what extent companies use this opportunity.
The test will be performed by analyzing the 10-K report of 50 different companies. This has been done on companies in S&P 500 that have been randomly selected in a period from 2008 until 2017, with an equal amount each year. The results of this test are shown in Table 5.1.

Table 5.1: Number of companies that include a part of the 10-K report by reference

Filing year	No reference	Reference in part 1	Reference in part 2	Reference in part 3	Reference in part 2 and 3
2008				5	
2009				5	
2010				4	1
2011				4	1
2012				4	1
2013				5	
2014				5	
2015				4	1
2016	1			4	
2017				4	1
Total	**1**	**0**		**39**	**5**

The categories in Table 5.1 are: no reference, reference in part 1, reference in part 2, reference in part 3 and reference in part 2 and 3. In these categories, it is possible to use references for single items, but the categories do not take this into account.
In Table 5.1, it is shown that companies in almost all cases refer to the proxy statement when incorporating part 3, which is fairly consistent between years as well. The effect of this is assumed to be minor, since it is mostly about corporate governance, and it is therefore not in this part the company will show any opinions regarding the future of the company.

A bigger problem, on the other hand, is that the test shows that 5 different companies refer to the annual report when incorporating part 2. This section describes how the company has performed and their outlook for the future, which makes it one of the most important parts of the 10-K when the sentiment of is analyzed. It is, therefore, something that has to be considered when the results are being analyzed.

5.2 Sample Selection

In the previous sections, the different challenges that were faced by using the 10-K report were described. This section, on the other hand, will describe how these 10-K reports will be combined to make the final data sample to base the analysis on. The section presents the different choices that have been made such as the sample period, which stocks that should be included, and how these are merged with the 10-K reports. The aim of this section is to explain the different choices that led to the final construction of the data sample.

5.2.1 Sample Period

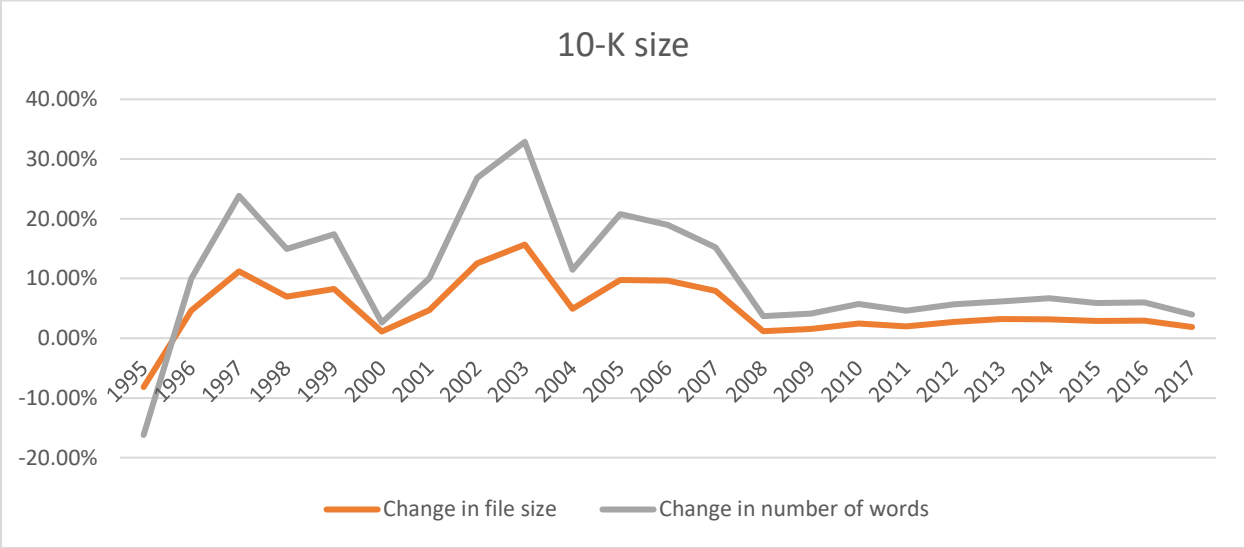
During the years there have been multiple changes to the disclosure requirements of 10-K reports. Regulators require more information and the length of the report will therefore increase. It is, therefore, necessary to make sure that data is consistent between years, and the number of positive and negative words in the 10-K reports will not increase due to the disclosure requirements changing significantly. The file size and word count of the 10-K reports have therefore been evaluated. This evaluation has been performed on 105.824 10-K reports on public companies from 1994 until 2017.

The result of this test is shown in Figure 5.1 below. As shown, there are some major spikes over the years. An example is in 2003, which is the effect of the implementation of the Sarbanes-Oxley Act in 2002. The purpose of the Sarbanes-Oxley Act was to improve the precision and accountability of the information provided by companies which resulted in the increase of disclosure requirements.

From 2008 until today the increase in file size has, however, stabilized, and there have been no major jumps in file size. This stable development, therefore, indicates that there have been no major changes in the disclosure requirements.

On top of making sure that the quality of data is not affected by the year of the report, it is also important to evaluate the results on an entire business cycle. The reason is that over time companies' growth, risk and profitability may be affected differently across the business cycle (Petersen & Plenborg, 2012, p. 66). The sentiment between companies will therefore presumably vary depending on the stage of the business cycle as well. To make sure that both up- and downturns in the economy is incorporated in the analysis, reports from 2008 until today will be used. The reason is that in 2007 initial signs of a financial crisis started to show and in late 2008 it developed into a full-blown crisis when Lehman Brothers filed for bankruptcy (Federal Reserve Bank of ST. Louis, 2018). This is also shown in the GDP of the USA, which for the first time since 1960 fell between 2008 and 2009. Since 2009 and until today the GDP of the USA has, however, increased every year (Appendix 2). Hence, by taking 2008 as the starting point both up- and downturns will be incorporated in the analysis. The amount of data based on years is therefore restricted to 10 years, and only reports from 2008 to 2017 will be included.

Figure 5.1: Change in size of 10-K reports



5.2.2 Evaluation of S&P 500

The S&P 500 is an index consisting of 500 U.S. companies in the large capitalization segment, which have common stock listed on U.S. stock exchanges. The index measures the performance of these large-cap companies, which captures approximately 80% of the available market capitalization in USA (S&P Dow Jones Indices, 2018). The index is therefore considered to be a proxy of the U.S. equity market. (S&P Global, 2018, p. 3). The reason why the companies in the S&P 500 index have been chosen to base the thesis upon will be described in Section 5.2.2.1 and in Section 5.1.2.2 it will be described how companies that are deleted from the index will be handled.

5.2.2.1 The S&P 500 Index

The composition of the S&P 500 index was developed and is still maintained by S&P Dow Jones Indices. To be considered as eligible for the index, companies must meet different criteria, which are rules regarding:

- The company's domicile
- The list of exchange the company is trading at
- The company's organizational structure and share type,
- If the company have multiple share classes
- The market capitalization of the company
- The liquidity of the stock
- The stocks investable weight factor (IWF)
- If the company is financially viable
- The amount of time it takes before a company can be included after an IPO

The above-mentioned criteria state, among other things, that to be eligible for the S&P500 index a company has to file an annual 10-K report, the primary listing has to be in the U.S., the market capitalization has to be above $6.1 billion and the companies have to deliver positive earnings in the sum of the most recent four consecutive quarters. S&P Dow Jones Indices, who oversees the making of the index, believes that turnover in the index membership should be avoided. The eligibility criterion is therefore primarily used as an addition criterion. Stocks that temporarily violate one or more of the addition criterions will therefore not be deleted from the index. Companies will only be deleted from the index if it substantially violates one or more of the eligibility criteria or is a part of a merger, acquisition or significant restructuring (S&P Global, 2018, pp. 5-8+24).

Based on the considerations mentioned above, the companies in the index are viewed as suitable to base the thesis upon. The reasons are especially that all companies in the index are required to file 10-K reports, the index is a good proxy of the U.S. equity market, and the changes in the index is kept at a minimum.

5.2.2.2 Survivorship Bias

In finance, it is quite common that retrospective studies are subject to what is defined as "survivorship bias". Survivorship bias describes the effect of excluding input that did not make it past some selection criteria, which will, therefore, make the results biased (McLeish, 2005, p. 236).

In the context of using S&P500, the survivorship bias may arise, if companies are included and afterward deleted from the index. This can happen if companies, as described in Section 5.2.2.1, substantially violate one or more of the eligibility criteria. In the analysis, companies that are deleted from the index but have been in the index between the sample period of 2008 to 2017 will therefore not be deleted from the dataset. The effect is that even though the companies are no longer in the index, they will still be included in the analysis which to some degree will prevent the survivorship bias.

Because of the abovementioned choices, the analysis will be comprised of companies whose stocks have been added to the S&P500 index between 2008 and 2017. The list of index additions is obtained from Compustat's list of index constituents (WRDS, 2018a). During the sample period from 2008 to 2017, there were 731 companies that were either already present or added to the index. All these companies will, due to the survivorship bias, be included from the year they are added, even though they may be deleted afterward.

As a starting point, the above-mentioned companies will be included, but there might be reasons that companies will get deleted from the analysis anyway. This might happen if the companies are part of a merger, acquisition or they have defaulted. An example of this is the merger of "The Dow Chemical Company" (Dow) and E.I. du Pont de Nemours & Company (DuPont), which merged into DowDuPont on August 31, 2017 (Dow Global, 2017). When these two companies merged they continued in a new company. The effect of this was that they got a new CIK (Central Index Key) code.

A CIK code is an identifier in SEC's computer systems to identify corporations and individual people, who have file disclosures with SEC (U.S. Securities and Exchange Commission, 2017). These CIK codes are therefore used to match the 10-K reports with the different stock returns. The effect of this is that if the companies for some reason, such as a merger, acquisition, or default change their CIK code, then it is this CIK code of the continuing company that will be used as an identifier. In the example of DowDuPont, the effect is therefore that the merged company should be included, but it is not possible since there are no 10-K reports available before 2018. Instead, the two separate companies of Dow and DuPont will be included on a separate basis until their merger in 2017.

5.2.3 Change in Sentiment Compared to Levels

The sentiment is measured by the proportion of positive and negative words or sentences out of the total amount of words or sentences in the 10-K report. When analyzing on these sentiment scores, it can be based upon levels or changes. The difference between these two is that an analysis based on levels is advantageous when the sentiment of separate companies is compared, while an analysis using changes in the sentiment score would be a better fit for a comparison of sentiment between years (Loughran & McDonald, 2016, p. 1219).

Instead of using the level of positive or negative sentences, the analysis will instead be based on the change from the previous year. The reasons for this choice are:

1. When an investor is using a 10-K, it can be argued that the point of interest is new information. For instance, the MD&A part of the 10-K will be uninformative if the firm does not change its MD&A disclosure notably when a significant economic change has happened (Brown & Tucker, 2011, p. 309+315).
2. The measure of change in tone rather than levels is consistent with prior research. It has furthermore in prior research been shown that the autocorrelation of tone levels between different periods are high. The correlation has shown to be around 65% and 70% when comparing the current and prior period (Feldman, Givindaraj, Livnat, & Segal, 2009, pp. 926-927). It has furthermore been shown that there is a lot of boilerplate information present in the report, meaning there is minimal change in the report. This has the effect that there will be minimal changes in the tone levels as well (Clarkson, Kao, & Richardson, 1999, p. 117+119).
3. The third reason is that positive and negative words or sentences in some cases will be specific to a company or industry. When a cross-sectional analysis is made where companies across different industries are included, this effect will have to be taken into account. When changes in sentiment are used, this problem will be minimized, because the usage of the same positive and negative words will presumably be stable over time in an industry or for a specific company. Changes in positive or negative sentiment will, therefore, be more robust (Feldman, Givindaraj, Livnat, & Segal, 2009, pp. 926-927).

As described above, changes in sentiment instead of levels will be used in the analysis. The effect on the dataset will be that the number of observations will be reduced. The reason is that the 10-K reports must be for two consecutive years to be included. An example of this is that there will be no observations in 2008 since this is the starting point. Thus, the first observations will be in 2009.

Before using changes in sentiment there must be made a final adjustment to the sample, which is to eliminate observations, where the reports consist of less than 100 words. The reason is that reports of this size have little information content, and when changes in sentiment are examined, it might influence the results considerably. This will only be the case for some reports filed in 2017, meaning the effect on the sample size will be minor.

5.2.4 Merging of Stock Returns and 10-K Reports

In the previous parts of Section 5.2, the number of years and the companies that are included in the analysis have been described, which was based on different criteria related to the 10-K reports and the S&P 500 index. This section will describe the process of joining the 10-K reports with their respective stock returns. These 10-K reports and the market information related to the companies represented in the S&P500 index is derived from different databases, and they are therefore not easily merged. The reason is that every company has a unique identifier, but since the 10-K reports and the market information is retrieved from different databases, these identifiers are not identical.

The 10-K reports have a unique identifier in their CIK code, and the same is the case with the constituent list retrieved from Compustat, and these two types of data are therefore easily merged. In CRSP, however, companies are not identified by their CIK code, but with a PERMNO code. PERMNO is a unique permanent security identification number assigned by CRSP to each security, and unlike other identification codes such as CUSIP, Ticker Symbol or Company name, the PERMNO does not change during an issue's trading history or after it is resigned after an issue (CRSP, 2018c). This identification code has been used, since the code is specific to a single security, and since companies may have multiple stock classes, it is therefore important that the identifier is based on a security level, rather than company level.

To merge the stock data and the 10-K reports CRSP has made a database available that merges CRSP and Compustat data (WRDS, 2018b). This database has been used to merge the PERMNO from CRSP and the CIK code from Compustat, which in the end makes it available to merge the different databases.

The approach was, therefore, to retrieve a dataset with all PERMNO on a daily basis, based on the CIK codes included in the S&P 500 constituent list. By doing this every PERMNO was matched to a CIK code. After this list was retrieved there were made different quality checks of the data to make sure the data was matched correctly. The most important check was to make sure that a PERMNO or CIK was not represented multiple times. If this was the case, the PERMNO or CIK were examined more deeply in the CRSP database to make sure the PERMNO or CIK number were the ones that should be used, and data were merged in the correct way.

In the end, the different PERMNOs could be used to merge the different stock returns in the CRSP database to a CIK code in the Compustat database. When this has been done all data is matched to a CIK number, making it possible to merge the stock returns to the different 10-K reports as well. The merged dataset will, therefore, consist of the observations created from the 10-K reports and the stock returns.

5.2.5 Stock Returns

The stock returns that the analysis will be based upon have been retrieved from CRSP on the different companies included in the S&P 500 index. These returns are based on the holding period return, which shows the total return by holding the stock over some period of time. When assessing the return of an investment, it is important to not only consider the price of the stock, but also other factors such as splits and dividend payments, which will affect the total return of the investment. The holding period return is therefore based on the total change in value of the stock over the time period. The formula for calculating the holding period return is:

$$r_t = \left[\frac{p_t * f_t + d_t}{p_{t-1}}\right] - 1$$

(CRSP, 2018b)

Where the variables are defined as:

- p_t = Price at time t.
- p_{t-1} = Price at last trading day before time t.
- f_t = Price adjustment factor that adjusts the stock prices after a distribution, such as dividends and splits.
- d_t = Cash adjustments at time period t, which can be dividends payed in cash.
- r_t = Holding period return from time t-1 until time t

5.2.6 Window-Size

When an annual report is filed, the textual parts will not necessarily be incorporated in a short period of time, as described in Section 2.2.1. To test if the BOW and RNTN model are able to quantify the textual parts of 10-K reports different window-sizes will be used. This setup is made to examine if it is possible to get predictive content out of the textual parts of 10-K reports, and at what time period the stocks reflect this content.

In this context, the window size describes the holding period return, which is the return of a specific period of time. Different window-sizes will be tested, but the initial test will be based on two trading days, which is defined as $[-1, +1]$, with day 0 as the filing date with SEC. This definition of window-size has been used in previous research

as well (Feldman, Givindaraj, Livnat, & Segal, 2009, p. 933), but the main reason is that there is no information available that defines the time of the day an annual report is filed with the SEC. If the report is filed after the market is closed, the information will not be incorporated in the price if the window size was $[-1, 0]$. Due to precaution, a window size of two trading days is therefore chosen since this makes sure that it has been possible to incorporate the information contained in the annual report in the stock price to some degree. By using one day prior to the filing date it is furthermore ascertained that any information leakage that might affect stock prices is captured, which is consistent with other studies (Feldman, Givindaraj, Livnat, & Segal, 2009, p. 933).

Other window sizes of longer time horizons than two trading days will also be evaluated. These tests will be made since new information is not necessarily incorporated in the price immediately. As described in Section 2.2.1, a post-earnings announcement drift is especially present up to 60 trading days after an earnings announcement. The tests will, therefore, be made on window-sizes of up to 60 trading days after the 10-K report has been filed, to test if any late reactions to the sentiment scores are shown.
These long window-sizes will take the post-earnings announcement drifts into account. By using this approach, it will first of all be tested if the sentiment scores of 10-K reports are able to explain the changes in stock price. Second, the effect of the window size will be tested, which will examine whether the information in the 10-K report is exposed to a time-lag.

The setup of window-sizes will be based on the holding period returns, which makes it easy to vary the different window sizes. The reason is that since the holding period returns have been calculated for one trading day, the window sizes are increased easily by the following formula, where a 2 days return has been calculated as an example:

$$(1 + r_t) * (1 + r_{t+1}) - 1 = 2\ days\ return$$

When different window-sizes are used it will have an effect on the number of observations that are used in the analysis as well. The reason is that there must be stock prices available at all trading days in the different window-sizes, otherwise, the observation will not be taken into account, and not be included in the final sample. One effect of this restriction is that if a company in a 60-day window-size has any missing stock returns, then all observations regarding that filing will be eliminated from the sample. The reason is that even though it might be possible to make the analysis on a window-size of two days, it will not be consistent to do this since the different regressions on window-size will be based on different samples of data and therefore not comparable.

5.3 Final Sample Selection

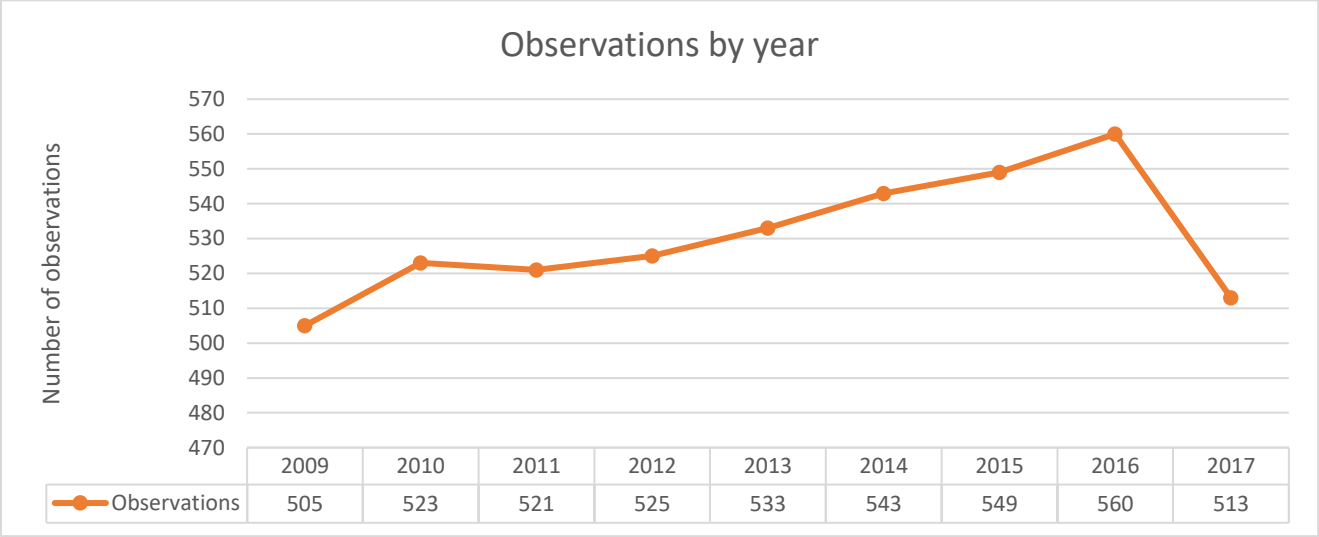
In the previous sections, the reasons behind the final choice of data sample have been described. The effect of these different choices has been compressed into Table 5.2. This table shows the effect of the different choices that have been made on the number of observations and companies which are included in the sample. As it is shown, the amount of observations has decreased considerably, before the final sample was selected. In the beginning, the dataset consisted of 294.325 observations related to 40.264 different companies. In the end, this sample has decreased to 4.772 observations from 669 different companies, which was the final sample that the analysis is based upon. The final sample selection can furthermore be grouped into the number of observations per year. This grouping by year is shown in Figure 5.2.

It is furthermore evident that the number of observations has an upward going trend and a drop in 2017. These findings are to be expected since in this dataset a company that has been included in the index is still included in the analysis even though the company is deleted from the index in the following year. The drop in 2017 is expected as well. This drop in observations is caused by the selection criteria that all observations need to have a stock price in 60 trading days preceding the filing of the 10-K report. Companies that file their 10-K report in one of the last months of 2017 will not have available stock data 60 days forward, since CRSP, which is the database where the stock data is retrieved from, does not offer any stock data beyond 2017.

Table 5.2: Sample Selection

Selection criteria	**Number of observations**	**Number of companies**
All annual filings of 10-K, 10-KT and 10K/A reports from 1994 to 2017.	294.325	40.264
Only 10-K and 10-KT reports are used.	176.410	34.519
Sample period ranges from 2008 until 2017.	84.228	16.568
Companies must be in the S&P 500 index between 2008 and 2017, but they are still included in the sample if they have been in the S&P 500 index and subsequently deleted from it.	5.552	718
There must be a stock price on all trading days up to 60 days after the filing of a 10-K report.	5.335	706
There must be 10-K reports for two consecutive years.	4.779	669
All observations where the 10-K report consists of less than 100 words are eliminated.	4.772	669

Figure 5.2: Number of observations by year



5.4 Regression Construction and Evaluation

This section will describe the method of uncovering the relationship between the BoW and RNTN model's sentiment scores of 10-K narratives and the subsequent stock returns. The statistical significance of the initial results will subsequently be tested by making different robustness checks. The robustness checks will begin by adjusting the stock returns for their inherent systematic risk using the Fama-French three-factor model, and afterward, different control variables will be added to the regression. Furthermore, an investigation on the effect of outliers in the risk-adjusted stock returns is presented.

To examine if there is a connection between the sentiment scores and the stock returns, a linear regression is used. The stock returns, which are the dependent variable, are regressed on the positive, negative and net sentiment score. This approach is used at the different window-sizes, which were described in Section 5.2.6. This approach is used as an initial method to uncover whether the sentiment scores are able to explain the changes in stock price before adjusting for systematic risk and including control variables.

To make these regressions, SAS – JMP is used. SAS – JMP is a statistical software, which provides cutting edge and modern statistical methods. It is furthermore a point and click software, which makes it easy to do advanced statistical analysis of data.

5.5 Dependent Variable

In Section 5.2.5 it was described how the raw stock returns were calculated. These stock returns will initially be adjusted with regard to the risk-free interest rate and afterwards for systematic risk as well.

To adjust the stock returns with regard to the risk-free interest rate and systematic risk, Kenneth R. French has made

a dataset available that can be used for this purpose. This dataset is based on daily observations, which therefore matches the holding period returns. The variables in the dataset are the one-month U.S. Treasury Bill rate and the returns on the three different Fama-French factors. (French, 2018).

5.5.1 Excess Returns

The first adjustment that will be made to the raw stock returns, is to subtract the risk-free interest rate from the holding period stock returns. The one-month U.S. Treasury Bill rate from the Kenneth R. French dataset will be used as the risk-free interest rate. By subtracting the risk-free interest rate, the dependent variable will become excess returns. When excess returns are used instead of the raw holding period returns, it is ascertained that fluctuations in the risk-free interest rate across different years, do not have any effect on the dependent variable.

The choice of the U.S. one-month Treasury Bill rate as the risk-free interest rate can be discussed. It has however been assessed to be a good choice due to the fact that the risk-free interest the analysis is based on the U.S. equity market and that government bonds are arguably the only securities that have a chance of being risk-free (Damodaran, Damodaran Online, 2018a).

5.5.2 Risk-Adjusted Returns

When the impact of the information content of 10-K reports is examined it is important to adjust the stock returns for systematic risk. The reason is that the model might otherwise explain the part of the stock returns that are due to the systematic risk and not due to the information content of the 10-K report (Lev & Ohlson, 1982, pp. 287-288).

The adjustment of risk is done by using the three-factor Fama-French model. The reasons for using the three-factor Fama-French model is described in Section 2.2.2 and especially concerns the lacking validity of the CAPM. As described in Section 2.2.2 the three-factor Fama-French model consists of the average return of small minus big portfolios (SMB), high minus low portfolios (HML), and lastly the excess return on the market, which have all been retrieved from the Kenneth R. French dataset described in the beginning of Section 5.5. These portfolios of returns include all stocks on the NYSE, AMEX and NASDAQ exchanges. The returns of the three factors in the Fama-French three-factor model will be regressed simultaneously on the daily excess returns for each stock, which is done to find the beta values (loading factors) of the three Fama-French factors. When these beta values are found it is possible to find the inherent systematic risk for the different stocks, which is found in the following way:

$$Systematic\ risk = \beta_i * R_t + s_i * SMB_t + h_i * HML_t$$

(Bodie, Kane, & Marcus, 2014, p. 340)

In the formula above the different betas has the notation i which represents the specific company, and the different variables the represents the returns has the notation t which represents the daily return on the market index, small

minus big portfolio or high minus low portfolio.
This systematic risk is subtracted from the excess return calculated earlier, and the returns of the different stocks will therefore be adjusted of systematic risk.

5.5.3 Beta Calculation

In Section 5.5.2 it was described that the beta values (loading factors) were calculated to adjust the returns for systematic risk. When betas are estimated there are particularly three issues that can have an influence on the beta calculation. The issues are the choice of market index, the time period used and the return interval which will be described below.

Choice of Market Index

There are in practice no indices that come close to measure the market portfolio. Instead, there are equity indices that measure the returns on a subset of securities in each market, which can be used as a proxy for the market portfolio. To use a combination of the stocks included on NYSE, AMEX and NASDAQ exchanges, therefore, seems like a reasonable assumption. The reason is that a better estimation of beta values will be made when more securities are included in the index, and furthermore, by using the securities included in these exchanges, it is ascertained that the index is diversified. The index is furthermore calculated by value weighting the returns, which takes the size of the different companies into account. Finally, it is reasonable to only use securities on these exchanges, since the companies the analysis is based upon are U.S. companies, and the securities included on these exchanges are primarily American (Damodaran, Damodaran Online, 2018b, pp. 6-7).

Time Period

The time horizon that the different beta values should be based upon is not specified in the Fama-French three-factor model. In the estimation of the different company betas, a time horizon of 10 years has been used. The 10-year period ranges from the beginning of the sample period in 2008 to the end in 2017. By going this far back in time the advantage is that the beta will be based on more observations, but on the other hand, it assumes that the characteristics of the different companies stay the same (Damodaran, Damodaran Online, 2018b, p. 8). The assumption that the characteristic of the companies stays the same might be a questionable assumption, but since it is only companies in the S&P 500 index that are examined, the companies the regression is based upon are relatively large and due to the S&P 500 eligibility criteria, they will also be relatively stable, and it can, therefore, be argued that a period of 10 years can be used to calculate the different beta values.

Return Interval

The final choice that will affect the beta values is the return interval, where daily returns have been chosen. Using short intervals increases the number of observations in the regression, which will improve the validity of the beta estimates. When a short return interval, such as daily returns, is used it will affect the beta estimates if there are periods where the stocks are non-trading during a return period (Damodaran, Damodaran Online, 2018b, pp. 9-10). Since the stocks in the S&P 500 has high liquidity in their stocks, the problem of non-trading is not a problem, and it, therefore, seems reasonable to calculate betas based on daily returns.

5.6 Adjusted R^2

In the previous section, it was described how the initial regressions will be applied. To examine which models perform the best, a metric must be used. The metric that will be used is the adjusted R^2.

The interpretation of R^2 is the percentage of the dependent variable that is explained by the independent variables in a linear regression, and it therefore seems easy to compare different models and choose the one with highest R^2. The effect of this is that when more independent variables are added to the regression the R^2 will therefore in almost all cases be higher . The effect of this is that instead of making the model better at explaining the dependent model, it might instead overfit the data. To take this into account adjusted R^2 is used to compare the different models, since it will take the number of variables into account (Newbold, Carlson, & Thorne, 2013, p. 492).

Another factor to take into account when comparing models based on adjusted R^2 is the input that goes into the regression to make it possible to compare. The general problem is that except for linear models with an intercept the R^2 is not comparable. The reason is that unless if it can be justified in data that no intercept should be included it will give misinterpretations of R^2. Furthermore, the models that is compared have to be linear, since transformations of variables will make the regression non-linear, and the it will therefore not be possible to compare the R^2 of this regression (Kvålseth, 1985, pp. 279-281).

All regressions in the analysis will therefore include an intercept term, the variables will not be transformed, and it will therefore be possible to compare the adjusted R^2 of the different regressions.

5.7 Control Variables

When the information content of using sentiment scores to explain the variations in stock returns is to be explained, it is important to control for other variables that might explain the variations in stock returns.

When the effect of the sentiment score is examined, it is important to make sure that that sentiment score is not capturing effects that are explained by other variables. It is particularly important if these other variables are included in the 10-K report, and therefore affects the share price over the filing date as well. These variables especially

concern different financial variables, that are included in the income statement or balance sheet. Prior research has taken this into consideration, some researchers use accruals (Feldman, Givindaraj, Livnat, & Segal, 2009, p. 927), while others use net income (Mitra & Hossain, 2009, p. 286). These measures are based on relative sizes of the companies since accruals or net income are divided by the size of the company which is measured by total assets or the market value of equity (Mitra & Hossain, 2009, p. 285) (Feldman, Givindaraj, Livnat, & Segal, 2009, p. 934). In this thesis, a variation of net income will be used. The usage of net income will be based on two variables, the levels of net income, and the change in net income compared to the previous year. To be consistent with prior studies, it is the relative size of net income that will be used. The income measures will, therefore, be divided with the total assets of the company in the current year.

The reason for including both variables is that that the explanatory power of the model is expected to increase when both a transitory and permanent variable is used. It has furthermore been shown that the change in net income and the level of income can be used as a proxy for the market's perception of earnings quality (Ghosh & Moon, 2005, p. 590).

The second variable that will be included is a variable that takes into consideration if the company is in the S&P 500 index or not. As described in Section 5.2.2.2 companies who were in the index, and afterward were excluded from the index are still included. To take this effect into consideration, a dummy variable will, therefore, be included.

There are multiple other variables that could be included in the regression as well but including these variables might just overfit the dependent variable instead of adding any additional explanatory power.

In previous research, other variables have been taken into consideration as well. Examples of other variables include number of analysts, firm size, book to market ratio and industry. These variables have already been taken into consideration, which will be described next.

The number of analysts is not included even though it has been done in previous research. It has been described that the reasoning behind including this variable is that the signal in sentiment score would be more effective for companies that are less heavily followed (Feldman, Givindaraj, Livnat, & Segal, 2009, p. 927). Since this thesis only examines the change in sentiment of companies represented in S&P 500, and it is assumed that all of these are heavily followed, then this variable will only have a minor effect.

Two other factors that could be considered are firm size and book to market ratio. It is expected that smaller companies will have larger incremental information content when the change in sentiment is examined, since the information environment of smaller companies is expected to be worse, than it is for large companies (Feldman, Givindaraj, Livnat, & Segal, 2009, p. 918+945). The same is the case with value stocks that are more stable and presumably more understandable companies than growth stocks and the change in sentiment score will, therefore, be a weaker signal for these companies (Feldman, Givindaraj, Livnat, & Segal, 2009, p. 918+945). Instead of taking these factors into account by adding control variables, the returns have been adjusted. It was done when the Fama-

French three-factor model was used to adjust the excess returns of their systematic risk. If these factors were added, it would therefore not increase the explanatory power of the model.

The last factor that has been considered is industry, but since changes in sentiment are used instead of levels, this variable will have little to no explanatory power. The reason is that some choices of words might be specific to an industry or company, but when changes in sentiment are used, this problem is mitigated. The reason is that words that are used extensively in an industry will be relatively stable over time, and it will, therefore, have no effect on the change in sentiment score (Feldman, Givindaraj, Livnat, & Segal, 2009, pp. 926-927).

The variables that will be used as control variables will therefore be, the level of net income relative to the size of the company, the change in net income relative to the size of the company, and a dummy variable that describes if the company is in the S&P 500 index or not.

It might be argued that more variables could be included but based on the discussion of other possible control variables, it seems reasonable to use the three control variables that have been selected.

5.7.1 Final Regression

In the end, the final regression will consist of the dependent variable, which is the risk-adjusted excess stock returns, where the Fama-French three-factor model has been used, to make the adjustment. The different types of sentiment scores will be used as explanatory variables, which are based on the change in sentiment from the previous year. The three sentiment scores that will be used are positive, negative and the difference between these. To make sure that the explanatory power of the sentiment score on the risk-adjusted stock returns are not affected by other factors the two control variables will be included.

The linear regression model will, therefore, be the following:

$$R_i = \beta_0 + \beta_1 * NI_i + \beta_2 * \Delta NI_i + \beta_3 * IN\_SP500 + \beta_4 * Sentiment\_score_i + \epsilon_i$$

Where the variables are defined as follows:

R_i= Excess stock returns adjusted for systematic risk

NI_i= Levels of net income for the current fiscal year divided by total assets for the current year

ΔNI_i= Change in net income from previous year divided by total assets for the current year

IN_SP500_i= If the company is in the index the variable is equal to 1 otherwise it is zero

$Sentiment\_score_i$= Change in positive, negative or net sentiment score from previous year

ϵ_i= The error term

5.8 Outliers

In Section 6.2.2 the summary statistics showed that the stock returns are influenced by outliers. Outliers are defined as an observation that is far away from most of the other observations. These observations might have a large influence on the results, and it is, therefore, necessary to consider how to handle them before any conclusions are made based on the data sample (Ghosh & Vogt, 2012, p. 3455).

The outliers are handled in different ways depending on the reason why an observation is an outlier, which can be due to different errors ore that the observation is just an extreme value compared to other observations. In this thesis the problems with outliers especially concern observations with extreme values. These values are usually treated in one of the following three ways.

1. The outlier is kept in the dataset as it is, and is treated like all other observations
2. The observations can be winsorized, which will assign less weight or modify the observation, which will make it closer to the value of other observations.
3. The last possibility is to drop the variable from the sample

When one of these three approaches are used, there is a danger that poor estimates of the different coefficients are produced. If the observation is kept in the dataset the influence of the observation may be overvalued in the and not treated in any way there is a possibility that the influence of the outlier may be overvalued in the estimation of the coefficient. On the other hand, the use of the second and third option will have a possibility to undervalue the influence of the outlier. Even though this is the case winsorizing or eliminating observations has been the standard way to handle outliers (Ghosh & Vogt, 2012, pp. 3455-3456).

Winsorizing will, therefore, be used to handle outliers. When winsorizing is used a threshold is set. This threshold is normally set to 5%, which mean that all observations above the 95 percentiles will be replaced by the value of the 95^{th} percentile in the dataset. The same approach is used for observations below the 5^{th} percentile which will be replaced by the value of the 5^{th} percentile. The assumption behind this approach is that the outlier does not seem right, and the coefficients will there be improved if it looks like the other observations. The advantage of this approach is therefore that the value is not dropped even though it is an outlier, but it is still kept as the most extreme value in the dataset. The effect of the outlier is therefore decreased, and the bias, which a dropped value would create, is avoided (Ghosh & Vogt, 2012, p. 3456).

This approach will, therefore, be used to test if the results of the regression are consistent when the influence of outliers is decreased.

Chapter 6 – Results

This chapter contains the regressions run for the historical data on each of the sentiment scores. The results will be analyzed where the aim is to assess the significance of the RNTN and BoW model's sentiment scores of 10-K reports as predictors of stock returns based on various window-sizes. Initially, the choice of sentiment scores will be discussed, followed by a presentation of the descriptive statistics of the data before moving on to the results of the different regressions. Finally, based on the analysis, a conclusion will be made on whether the RNTN or BoW model is superior at predicting stock returns.

6.1 The Sentiment Scores as Explanatory Variable

The sentiment scores that will be used as predictors of stock returns in this analysis are the positive, negative, and net sentiment score. They will be tested on different window-sizes from 2 to 60 days and will be included in the models as explanatory variables on a separate basis.

The reason these sentiment scores have been chosen is to examine which sentiment score is superior to which model and whether their combined explanatory power has an effect. The initial estimate is that the positive sentiment score is superior for the RNTN, while the negative score is superior for the BoW approach. The reason for this is, as described in Section 4.3.1 that the RNTN has a bias when the negative score is examined, and the BoW model has a bias when the positive sentiment is examined. This bias is to some extent mitigated, however, by using the change in sentiment score from the previous year, which is also described in Section 5.2.3. Through this method, it will be possible to examine if the models perform as expected or whether one of the models has a higher explanatory power than the other, thus forming a basis for comparison.

In the regressions, a coefficient for each sentiment score will be estimated. This coefficient is a signal of how the stock returns are predicted to develop when the sentiment score changes. The interpretation of the different sentiment scores' coefficients is the following: if the positive sentiment score increases from the previous year ($\Delta Positive\ in\ \%$), then it is expected that the stock returns will increase, since this is estimated to be a signal of favorable information in the 10-K report. Therefore, the sign of the coefficient is expected to be positive. On the other hand, the change in negative sentiment from the previous year ($\Delta Negative\ in\ \%$) is expected to have a negative effect on stock returns, since this is estimated to be a signal of unfavorable information in the 10-K report. Hence, the sign of the coefficient is expected to be negative. The last coefficient is the net sentiment score, which is calculated by subtracting the change in negative sentiment from the change in positive sentiment

($\Delta Positive\ in\ \% - \Delta Negative\ in\ \%$). If the change in positive sentiment is larger than the change in negative sentiment it is expected that the stock returns will increase. Therefore, the expectation is that this coefficient will be positive. The expected signs of the different sentiment score coefficients are shown in Table 6.1.

Table 6.1: Expected signal of sentiment scores

Sentiment Score	**ΔPositive in %**	**ΔNegative in %**	**ΔPositive in %−ΔNegative in %**
Sign of coefficient	+	−	+

6.2 Descriptive Statistics

This section will present descriptive statistics for the different variables used in this thesis. The descriptive statistics will consist of the explanatory variables, including the applied control variables, the excess stock returns, the risk-adjusted returns, and the different sentiment scores.

6.2.1 Summary Statistics of the Explanatory Variables

Table 6.2 provides the summary statistics of the different explanatory variables.

Table 6.2: Summary statistics of the explanatory variables

	Mean	**Standard deviation**	**Minimum**	**5th percentile**	**Median**	**95th percentile**	**Maximum**
Negative Sentiment Score BoW	0,0386	0,3090	−5,1192	−0,3575	0,0277	0,4710	3,5185
Positive Sentiment Score BoW	0,0016	0,1045	−0,8158	−0,1516	0,0010	0,1517	1,6002
Net Sentiment Score BoW	−0,0370	0,3355	−3,4787	−0,5177	-0.0246	0,3967	5,3349
Negative Sentiment Score RNTN	0,0159	1,3293	−12,8608	−1,7980	-0,0171	1,7814	14,2085
Positive Sentiment Score RNTN	0,0369	0,7226	−6,3747	−1,0485	0,0302	1,1265	4,7257
Net Sentiment Score RNTN	0,0210	1,8558	−15,1576	−2,7328	0,0674	2,6476	12,2788
Level of income	0,0587	0,0817	−0,5505	−0,0436	0,0521	0,1729	0,5379
Change in income	0,0048	0,0826	−0,7781	−0,0864	0,0027	0,0976	2,0464
In S&P 500	0,9855	0,1194	0	1	1	1	1

In the summary statistics of the explanatory variables, it is evident that the variations in the sentiment scores are larger for the RNTN model compared to the BoW model. This is shown by the standard deviation, which is more than double the size compared to the BoW model.

Examining the means of the BoW model sentiment scores, it is evident that the negative sentiment score on average is larger than the positive sentiment score, which also makes the mean net sentiment score negative. For the RNTN model, on the other hand, the opposite is the case, which results in the mean of the net sentiment score becoming positive. It is surprising that the mean of the net sentiment score shows opposite results for the two models, which might be an initial indication that one of the models is better at extracting the sentiment of the narrative in the 10-K report than the other.

Regarding the three control variables, it is worth noticing that the positive median of the income variables shows that more than half of the reports that are included in the analysis show positive income, and they have improved this income from the previous year. In addition, the mean for the S&P 500 variable of 0,9855 shows, that only about 1,5% of the companies included in the analysis are not in the index at the filing date.

It can furthermore be described, that the sentiment scores are exposed to outliers since the maximum and minimum values deviate considerably from the 5^{th} and 95^{th} percentiles. One example is that the minimum value of the negative sentiment score for the BoW model is $-5{,}12$, while the 5^{th} percentile is only $-0{,}3575$. In the other end of the spectrum is the maximum value 3,5185, while the 95^{th} percentile is only 0,4710, which, therefore, is an indication that the sentiment scores are exposed to outliers.

6.2.2 Summary Statistics of the Excess Returns and Risk-Adjusted Returns

Table 6.3 shows the summary statistics on different window-sizes of the excess returns and Table 6.4 shows the Risk-adjusted returns.

Table 6.3: Summary statistics of excess returns

Window - Size	Mean	Standard deviation	Minimum	5th percentile	Median	95th percentile	Maximum
2 days	−0,0002	0,0569	−0,5122	−0,0654	0,0025	0,0499	2,5074
3 days	0,0006	0,0567	−0,5993	−0,0741	0,0036	0,0596	1,5253
4 days	0,0013	0,0645	−0,5564	−0,0838	0,0043	0,0682	1,8994
5 days	0,0015	0,0713	−0,5854	−0,0953	0,0054	0,0761	1,8526
10 days	0,0101	0,0833	−0,5932	−0,1018	0,0111	0,1086	2,0862
15 days	0,0215	0,0972	−0,6938	−0,0956	0,0175	0,1458	1,9996
30 days	0,0433	0,1260	−0,7219	−0,1053	0,0326	0,2262	2,3800
60 days	0,0621	0,2141	−0,7138	−0,1797	0,0410	0,3432	3,1553

Table 6.4: Summary statistics of risk-adjusted returns

Window - Size	Mean	Standard deviation	Minimum	5th percentile	Median	95th percentile	Maximum
2 days	0,0001	0,0504	−0,4061	−0,0413	−0,0002	0,0440	2,4682
3 days	0,0000	0,0478	−0,6105	−0,0483	−0,0001	0,0531	1,4907
4 days	0,0000	0,0529	−0,5799	−0,0549	−0,0006	0,0597	1,7035
5 days	−0,0001	0,0563	−0,5933	−0,0604	−0,0005	0,0636	1,6403
10 days	0,000	0,0704	−0,6185	−0,0862	−0,0007	0,0846	1,8398
15 days	0,0007	0,0812	−0,6573	−0,0986	−0,0001	0,0959	1,7752
30 days	0,0029	0,0938	−0,7275	−0,1247	0,0023	0,1215	1,2691
60 days	0,0099	0,1533	−0,7219	−0,1916	0,0058	0,2114	2,4834

When the summary statistics of the excess returns are examined it is evident that the mean return is around 0 at a window-size of 2, but as this window-size increases, the expected mean return increases as well. The reason for this is that these returns might be influenced by the systematic risk since they have not been adjusted for this yet. Following this logic, the risk-adjusted returns are expected to give a mean of 0 if the Fama-French factors fully explain the returns. In the table, it is evident that this is the case in the short window-sizes of 2 to 10 days, but when the window-size is increased to 15, 30 and 60 days, the mean becomes slightly larger than 0. Overall, the expected returns have decreased considerably after they have been adjusted for systematic risk.
It is furthermore evident from the summary statistics, that there might be outliers in the returns since the minimum and maximum returns for the different window-sizes are much higher than the returns of the 5th and 95th percentile. An example of this is that the maximum excess return at a window-size of 2 days is 250,74%, while the observation at the 95^{th} percentile is only 4,99%. The same apply to the minimum observation where the excess return is $-51{,}22\%$ while the 5^{th} percentile is −6,54%. These outliers apply to all window-sizes, and the same is the case if the risk-adjusted returns are examined.

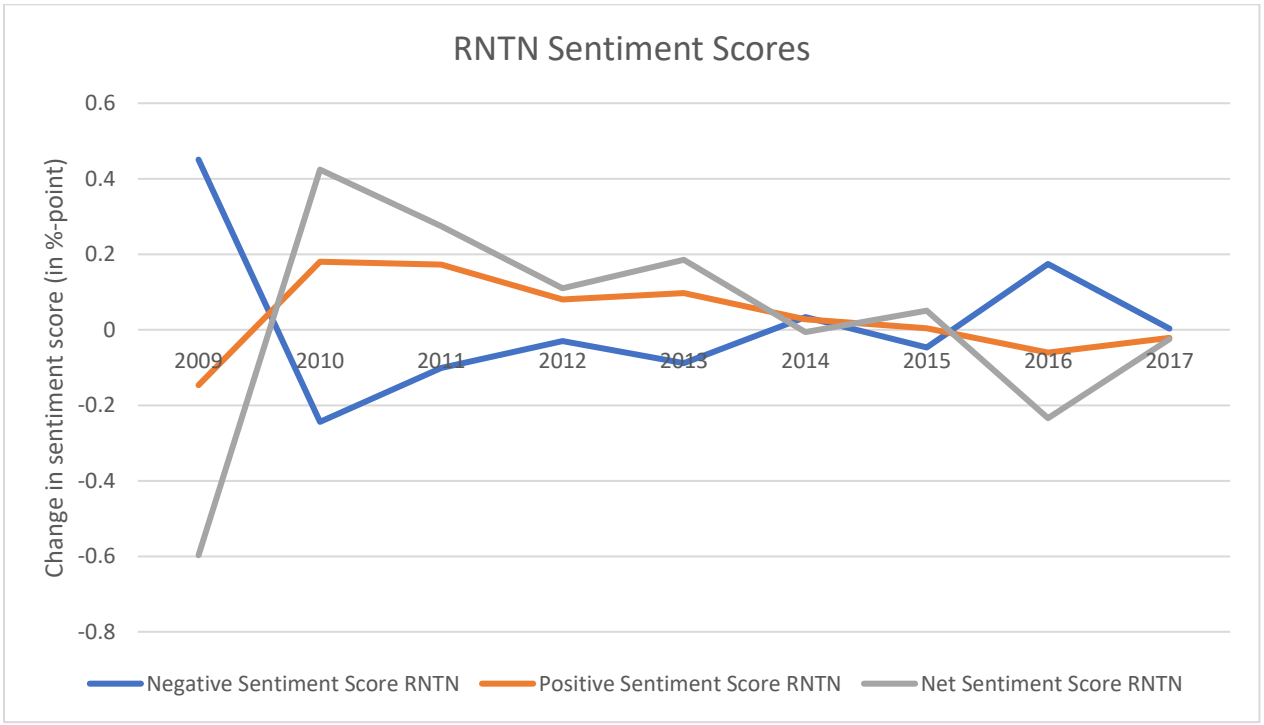
6.2.3 Development of Average Sentiment Scores

In Figure 6.1 and 6.2, the development of the average change in sentiment score from the previous year is shown. The input that is used to make the graphs are shown in Appendix 5.

Figure 6.1 and 6.2 show that the change in the different sentiment scores for both models is fairly consistent over the years since the sentiment scores are around 0 in most years with only minor fluctuations. These fluctuations are more noteworthy for the RNTN though. The only exception is in 2009, where a major change is recorded, which is the case for both models.

Figure 6.1: Change in sentiment scores for the BoW model

Figure 6.2: Change in sentiment scores for the RNTN model



6.2.4 Correlation Matrix

The table below provides a correlation matrix between the sentiment scores and the control variables, which are used as explanatory variables.

In this matrix, it is particularly interesting to examine if the sentiment scores from the RNTN are correlated with each other. This applies to the BoW model's sentiment scores as well.

The other factor that is interesting to examine is if any of the control variables are highly correlated with the sentiment scores or each other. Said in another way, it is interesting to examine if the variables are exposed to multicollinearity.

It is furthermore possible to test if the correlation coefficients are significantly different from zero. It will, therefore, be examined if there is no linear relationship between the variables, which will be done at a 5% significance level. This will be tested using the following two-sided decision rule:

$$-t_{n-2,\frac{\alpha}{2}} > \frac{r * \sqrt{n-2}}{\sqrt{(1-r^2)}} > t_{n-2,\frac{\alpha}{2}}$$

(Newbold, Carlson, & Thorne, 2013, p. 453)

- $\alpha =$ Significance level, which is defined as 5%
- $t =$ Student's t Distribution with 4772-2 Degrees of Freedom. The critical value is therefore 1,960
- $r =$ The sample correlation
- $\rho =$ The population correlation

If the decision rule above is satisfied, the H_0 hypothesis that the population correlation is equal to 0 will be rejected. This will happen, if $r = \pm 0{,}0284$.

If the H_0 hypothesis is rejected, the correlations are marked with an asterisk in the correlation matrix below (Table 6.5).

The correlation matrix shows that the correlations between the net sentiment scores are significantly correlated with the other sentiment scores for both models, which is to be expected. This is especially the case of the negative sentiment score, where the correlation with the net sentiment score is $-0{,}9504$ for the RNTN model and $-0{,}9507$ for the BoW model.

When the correlations of the control variables with the sentiment scores are examined, it is evident that the correlations are low, and the correlations do therefore not show any sign of multicollinearity. This sign becomes even stronger when the significance of these correlations is examined since it cannot be rejected that the true correlation between the sentiment scores of the RNTN model and most of the control variables are actually 0 at a

5% significance level. Therefore, it is valid to use these control variables because their lack of multicollinearity will not give misleading results.

Table 6.5: Correlation matrix of explanatory variables

	Positive Sentiment Score RNTN	Negative Sentiment Score RNTN	Net Sentiment Score RNTN	Positive Sentiment Score BoW	Negative Sentiment Score BoW	Net Sentiment Score BoW	Level of income	Change in income
Negative Sentiment Score RNTN	−0,6012*							
Net Sentiment Score RNTN	0,8200*	−0,9504*						
Positive Sentiment Score BoW	0,2550*	−0,1003*	0,1712*					
Negative Sentiment Score BoW	−0,1497*	0,1789*	−0,1864*	−0,0955*				
Net Sentiment Score BoW	0,2173*	−0,1960*	0,2250*	0,3994*	-0,9507*			
Level of income	0,0159	−0,0110	0,0140	0,0284	−0,0895*	0,0912*		
Change in income	0,0256	−0,0403*	0,0388*	0,0625*	−0,1019*	0,1133*	0,3626*	
In S&P 500	0,0071	−0,0264	0,0213	−0,0446*	0,0040	−0,0234	0,0964*	0,1242*

6.3 Preliminary Regressions

In this section, the preliminary regressions will be made, which will test whether the sentiment scores have any explanatory power on the excess returns before other explanatory factors have been included in the model. The following will, therefore, show the results of numerous regressions, where the three different sentiment score variables for each model have been included as the explanatory variable and regressed on excess stock returns of different window sizes.

If the different sentiment coefficients are significantly different from 0 at a significance level of 5%, they will be marked with an asterisk in the tables. This approach will be used throughout the rest of the chapter.

6.3.1 Negative Sentiment Score

Table 6.6 and Table 6.7 show the regressions of the change in negative sentiment on excess stock returns for the two models. The regressions have been performed on different window-sizes, which is shown in the left-most column.

Table 6.6: Regression results on negative sentiment score from BoW model

Window-size in days	Coefficient Negative Sentiment Score BoW	P-value Negative Sentiment Score BoW	Adjusted R^2
2	−0,0153	<0,0001*	0,0067
3	−0,0175	<0,0001*	0,0089
4	−0,0187	<0,0001*	0,0078
5	−0,0246	<0,0001*	0,0111
10	−0,0112	0,0043*	0,0015
15	−0,0022	0,6292	−0,0002
30	0,0332	<0,0001*	0,0064
60	0,0738	<0,0001*	0,0111

Table 6.7: Regression results on negative sentiment score from RNTN model

Window-size in days	Coefficient Negative Sentiment Score RNTN	P-value Negative Sentiment Score RNTN	Adjusted R^2
2	−0,0013	0,0423*	0,0007
3	−0,0017	0,0054*	0,0014
4	−0,0028	<0,0001*	0,0032
5	−0,0028	0,0002*	0,0026
10	−0,0027	0,0028*	0,0017
15	−0,0016	0,1259	0,0003
30	0,0028	0,0373*	0,0007
60	0,0093	<0,0001*	0,0032

When the results of the regressions are examined, it is obvious that both models have a very low adjusted R^2. The effect of such a low adjusted R^2 is that the change in negative sentiment only explains very little of the variation in the excess stock returns.

If the two models are compared based on the adjusted R^2, it is shown that the BoW model's adjusted R^2 is slightly higher, which can be due to the negativity bias of the RNTN model. Both models do, however, have low adjusted R^2 scores, which are below 1%, with only one exception, which is the 5 day window-size of the BoW model with an adjusted $R^2\ of$ 1,11%.

When the coefficients of the two models are examined, it is evident that the negative sentiment variable is significant at a 5% significance level at all window-sizes beside at 15 days, which is illustrated by the P-values. The interpretation of this is that it can be rejected on a 5% significance level that the coefficients equal zero. The negative sentiment scores are therefore able to explain some of the variation in the excess stock returns, since it can be rejected that the negative sentiment score has no effect. It is to some degree surprising that the RNTN is able to explain these variations since it has a negative bias, but since the yearly changes in sentiment scores has been used, some of this bias might have been eliminated, which makes the variable significant.
The sign of the coefficient of the negative sentiment scores for both models are consistent with expectations on the shorter window-sizes. This is however not true at the large window-sizes of 30 and 60 days, where the coefficients become the opposite sign for both models.

It is furthermore noteworthy that the size of the coefficients varies between the two models. If the 2-day window-size is used as an example, a one percentage point increase in the negative sentiment score is expected to give an expected negative excess return of $-1{,}53\%$ for the BoW model. The same increase in sentiment score will, on the other hand, give an expected negative excess return of $-0{,}13\%$ for the RNTN model. This difference in the size of coefficients shows that the BoW model is more sensitive to changes in the negative sentiment score, which is the case at the other window-sizes as well.

6.3.2 Positive Sentiment Score

Table 6.8 and Table 6.9 show the results of the regression of the positive sentiment score on the excess stock returns for the two models.

Table 6.8: Regression results on positive sentiment score from BoW model

Window-size in days	Coefficient Positive Sentiment Score BoW	P-value Positive Sentiment Score BoW	Adjusted R^2
2	0,0103	0,1941	0,0001
3	0,0155	0,0495*	0,0006
4	0,0153	0,0877	0,0004
5	0,0178	0,0713	0,0005
10	−0,0105	0,3629	0,0000
15	−0,0374	0,0056*	0,0014
30	−0,0870	<0,0001*	0,0050
60	−0,1604	<0,0001*	0,0059

Table 6.9: Regression results on positive sentiment score from RNTN model

Window-size in days	Coefficient Positive Sentiment Score RNTN	P-value Positive Sentiment Score RNTN	Adjusted R^2
2	0,0011	0,3251	0,0000
3	0,0018	0,1208	0,0003
4	0,0031	0,0170*	0,0010
5	0,0035	0,0135*	0,0011
10	0,0026	0,1130	0,0003
15	−0,0006	0,7512	−0,0002
30	−0,0070	0,0056*	0,0014
60	−0,0193	<0,0001*	0,0040

When the results of this regression are examined, it is noticeable that the adjusted R^2s are even lower than they were for the negative sentiment scores – for both models. The positive sentiment score is therefore worse at explaining the variations in excess stock returns. Regarding the BoW model, this is not a surprise because of its bias due to managements tendency of framing negative sentences with positive words in 10-K reports (Loughran & McDonald, 2016, p. 1217). It is, however, less clear why the adjusted R^2 is not higher for the positive sentiment scores for the RNTN model compared to the negative sentiment scores characterized by bias. It can be argued that by using changes in sentiment scores it mitigates this bias, which offers an explanation as to why the adjusted R^2 for the positive sentiment score is lower than for negative sentiment score.

If the signal and significance of the coefficients on the different window-sizes are examined, it can be described that it is only 4 out of 8 coefficients that are significant at a 5% significance level, which is the case for both models. It is especially on the short window-sizes that the coefficients are insignificant at a 5% level. The effect is therefore that it cannot be rejected that the coefficients are equal to zero. The variables that are significant particularly relate to the long window-sizes of 30 and 60 days, and as the case was when the negative sentiment scores were examined, the coefficients have an unintuitive sign. The coefficients of the 30 and 60-day window-sizes, therefore, describe that an increase in the positive sentiment score will have a negative effect on excess stock returns.
Furthermore, the size of the coefficients shows the same tendency as was shown in the negative sentiment score. It can, therefore, be described that the positive sentiment score of the BoW model is more sensitive to changes in the positive sentiment score compared the RNTN.
The results of this regression indicate that for both models the negative sentiment score is better at explaining the changes in excess stock returns compared to the positive sentiment score.

6.3.3 Net Sentiment Score

Table 6.10 and Table 6.11 show the results of the net sentiment score regressed on the excess returns for the two models.

Table 6.10: Regression results on net sentiment score from BoW model

Window-size in days	Coefficient Net Sentiment Score BoW	P-value Net Sentiment Score BoW	Adjusted R^2
2	0,0140	<0,0001*	0,0066
3	0,0164	<0,0001*	0,0092
4	0,0174	<0,0001*	0,0080
5	0,0226	<0,0001*	0,0111
10	0,0084	0,0190*	0,0009
15	−0,0018	0,6723	−0,0002
30	−0,0366	<0,0001*	0,0093
60	−0,0781	<0,0001*	0,0148

Table 6.11: Regression results on net sentiment score from RNTN model

Window-size in days	Coefficient Net Sentiment Score RNTN	P-value Net Sentiment Score RNTN	Adjusted R^2
2	0,0008	0,0655	0,0005
3	0,0011	0,0092*	0,0012
4	0,0019	0,0001*	0,0028
5	0,0020	0,0003*	0,0025
10	0,0018	0,0057*	0,0014
15	−0,0007	0,3280	0,0000
30	−0,0025	0,0101*	0,0012
60	−0,0077	<0,0001*	0,0043

When the results of the regression of the net sentiment scores on excess returns are analyzed, it is clear that the statistical significance of this variable is similar to the significance of the negative sentiment score variable. This is somewhat expected, since the correlation between the negative sentiment score, and the net sentiment score was −0,9504 for the RNTN model and −0,9507 for the BoW model. This correlation is reflected in the adjusted R^2 since the adjusted R^2 scores of the net- and negative sentiment scores are very similar. The similarities are also consistent with regard to the significance of the different coefficients. The results of the negative sentiment scores show that all variables for both models were significant at a 5% level, except for the coefficient at a window-size of 15 days. The same is the case with the net sentiment score, except that the coefficient for the window-size of two days is only significant at a 10% level, instead of on a 5% level.

The size of the coefficients is similar to the results of the negative sentiment score as well for both models which indicates again that the BoW model is more sensitive to changes in the net sentiment score than the RNTN model. The sign is, however, the opposite of the negative sentiment score. The reason is that a larger increase in the positive sentiment score compared to the increase in the negative sentiment score will expectedly give an increase in excess stock returns. This is reflected in the window-sizes of 2 to 10 days. On the other hand, the longer window-sizes give the opposite signal which was also the case with the negative and positive sentiment scores.

6.3.4 Explanation of Low Adjusted R^2

It was evident in the description of the preliminary regressions that the regressions experienced a low adjusted R^2. The fact that the variables are significant, but experience a low R^2 is consistent with Feldman's (2009, p. 918) research on stock returns in regard to management's tone change in the MD&A. His results showed among other things a R^2 of 0,16%. The variables included in Feldman's regression were a net sentiment score based on a dictionary approach and on top of that accruals were included as well. These variables were regressed on the excess return of a window-size of 3 days around the SEC filing date, which is therefore similar in many ways to the regression made with the BoW model's output in this thesis. If Feldman's results are compared to the regression results of the net sentiment score in this analysis, then it is evident that the adjusted R^2 in this analysis is higher in most cases for the BoW model. This means that the explanatory power can be argued to be higher. Furthermore, the coefficient is significant as the case was in Feldman's study.

Feldman explains that the contributions of using only qualitative information to explain stock returns are not large, which might explain the low R^2. The reason is that either the potential of qualitative information is low or the tools that are currently available for quantifying qualitative data are too inaccurate (Feldman, Givindaraj, Livnat, & Segal, 2009, p. 916). Based on the knowledge about the tools applied in this thesis, which are a good representation of the current state of the art models for textual analysis, it seems, all things being equal, that the latter might be the case. Even though this is the case the coefficients of the models are still significant, but it still makes sense that there is much information still waiting to be retrieved in the qualitative information.

6.3.5 Key Results

Overall, there are 6 key results from the initial regression of the three sentiment scores on the excess stock returns:

1. Overall the models are able to identify a trend in the stock returns on different window sizes, especially when the negative sentiment score and the net sentiment score is used as the explanatory variable, which is shown by the significance of the different coefficients. This is the case for both models.
2. The BoW model has a slight advantage compared to RNTN model, which is shown by a slightly higher adjusted R^2 score, which is especially evident for the negative and net sentiment score. The reason why the adjusted R^2 is slightly higher for the BoW model can be due to the negative bias of the RNTN model.

3. In all regressions, the adjusted R^2 is very low, which indicates that variations in excess stock returns are influenced by other factors.
4. In general, the sensitivity of the different coefficients is higher for the BoW model compared to the RNTN model. The effect is that an increase in sentiment score for the BoW model is expected to have a larger effect on excess stock returns than the RNTN.
5. The coefficients of both models have primarily the same sign, which is the case across the different sentiment scores and window-sizes.
6. The signal for both models are primarily as expected for the short window-sizes of 2 to 5 days, but when the window-sizes are increased to 30 and 60 days, the sign changes and becomes the opposite of what is expected.

These results provide a good foundation for further exploring the models' explanatory power of stock returns.

6.4 Risk-Adjusted Returns Based on Fama-French

In Section 6.3 regarding preliminary regressions, the sentiment scores were the explanatory variables, and the excess stock returns were used as dependent variable.
When the sentiment scores' impact on stock returns is examined, it is, as explained in Section 5.5.2, important to take the systematic risk of the returns into consideration. The reason is that the sentiment score might otherwise give misleading results. If the returns are not adjusted for systematic risk, then the explanatory variable might be explaining this systematic risk, instead of the incremental information content from the sentiment score.
The regression that will be performed will be similar to the ones in the previous section. The only difference is that the excess returns will be adjusted for their systematic risk. The adjustment of systematic risk is done by using the Fama-French three-factor model as described in Section 5.5.2.

The results of the regressions on risk-adjusted returns are presented in the same fashion as with the preliminary regressions. In addition, a comparison between the two will be given.

6.4.1 Negative Sentiment Score

Table 6.12 and 6.13 show the results of the negative sentiment score on the same window-sizes as in the preliminary regressions.

Table 6.12: Regression results on negative sentiment score from BoW model

Window-size in days	Coefficient Negative Sentiment Score BoW	P-value Negative Sentiment Score BoW	Adjusted R^2
2	−0,0059	0,0130*	0,0011
3	−0,0068	0,0022*	0,0018
4	−0,0053	0,0313*	0,0008
5	−0,0058	0,0288*	0,0008
10	−0,0022	0,4961	−0,0001
15	−0,0031	0,4095	0,0000
30	0,0088	0,0465*	0,0006
60	0,0272	<0,0001*	0,0028

Table 6.13: Regression results on negative sentiment score from RNTN model

Window-size in days	Coefficient Negative Sentiment Score RNTN	P-value Negative Sentiment Score RNTN	Adjusted R^2
2	−0,0001	0,8117	−0,0002
3	−0,0004	0,4962	−0,0001
4	−0,0007	0,2123	0,0001
5	−0,0005	0,4190	0,0000
10	−0,0007	0,3803	0,0000
15	−0,0003	0,7128	−0,0002
30	0,0011	0,2971	0,0000
60	0,0037	0,0259*	0,0008

Compared to the preliminary regression on negative sentiment, it is clear that the adjusted R^2 decreases. This reduction shows that some of the explanatory power in the preliminary regressions was due to systematic risk factors. It is, however, a bit surprising that the adjusted R^2 is almost 0 at all window-sizes for the RNTN model meaning it provides a marginal explanation of the variation in the risk adjusted stock returns over the filing date. Even though the adjusted R^2 has decreased for the BoW model as well, it still has some explanatory power.

The same tendencies are shown when the significance of the negative sentiment coefficients is examined. It is evident that except for the window-size of 60 days, the negative sentiment coefficients of the RNTN model have become insignificant at a 5% significance level, and it can therefore not be rejected that these coefficients are

different from 0. The implication of this is that it cannot be rejected that these coefficients do not explain any changes in stock returns over the filing date after the returns have been adjusted for systematic risk.

The coefficients of the BoW model, on the other hand, are significant at a 5% level except for the coefficients on window-sizes of 10 and 15 days. This gives support to the argument that the negative sentiment score of the BoW model can explain changes in stock returns over a filing date even though the returns have been risk-adjusted.

Finally, the level and sign of the negative sentiment coefficients are examined. Across all window-sizes, the coefficients have decreased compared to the excess returns. This makes sense given the adjustment of systematic risk generally causes the returns to be lower. The sign of the coefficient has, however, stayed the same for both models compared to the preliminary regressions on the negative sentiment score.

6.4.2 Positive Sentiment Score

Table 6.14 and Table 6.15 show the results of the positive sentiment score regressed on the risk-adjusted returns on the different window-sizes.

Table 6.14: Regression results on positive sentiment score from BoW model

Window-size in days	Coefficient Positive Sentiment Score BoW	P-value Positive Sentiment Score BoW	Adjusted R^2
2	0,0024	0,7334	−0,0002
3	0,0050	0,4454	0,0000
4	0,0033	0,6552	−0,0002
5	0,0035	0,6529	−0,0002
10	−0,0118	0,2283	0,0000
15	−0,0299	0,0078*	0,0013
30	−0,0409	0,0016*	0,0019
60	−0,0758	0,0004*	0,0025

Table 6.15: Regression results on positive sentiment score from RNTN model

Window-size in days	Coefficient Positive Sentiment Score RNTN	P-value Positive Sentiment Score RNTN	Adjusted R^2
2	0,0001	0,9064	−0,0002
3	0,0003	0,7396	−0,0002
4	0,0007	0,5075	−0,0001
5	0,0007	0,5440	−0,0001
10	0,0005	0,7224	−0,0002
15	−0,0010	0,5382	−0,0013
30	−0,0041	0,0309*	0,0008
60	−0,0087	0,0045*	0,0015

When the results of the regressions with the positive sentiment score as explanatory variable are examined, it is evident that the explanatory power of the positive sentiment score is low. This is the case for all window-sizes and for both models. The adjusted R^2 is around 0 for both models and the positive sentiment coefficients are not significant at a 5% level, when the window-size is at 10 days or below. When the window-size is at 30 or 60 days, both models show significant results, since the adjusted R^2 show some explanatory power, and the positive sentiment coefficients are significant at a 5% level. As it was the case in the preliminary regression on the positive sentiment score, the signal of the coefficient on the 30- and 60-day window-sizes, are not as expected. The ability of the positive sentiment score to predict the changes in stock returns across a filing data can therefore be questioned.

6.4.3 Net Sentiment Score

Table 6.16 and Table 6.17 show the results of the net sentiment score regressed on the risk-adjusted returns.

Table 6.16: Regression results on net sentiment score from BoW model

Window-size in days	Coefficient Net Sentiment Score BoW	P-value Net Sentiment Score BoW	Adjusted R^2
2	0,0052	0,0167*	0,0010
3	0,0063	0,0022*	0,0017
4	0,0048	0,0339*	0,0007
5	0,0052	0,0313*	0,0008
10	0,0008	0,8013	−0,0002
15	−0,0002	0,9448	−0,0002
30	−0,0114	0,0049*	0,0015
60	−0,0304	<0,0001*	0,0042

Table 6.17: Regression results on net sentiment score from RNTN model

Window-size in days	Coefficient Net Sentiment Score RNTN	P-value Net Sentiment Score RNTN	Adjusted R^2
2	0,0001	0,8282	−0,0002
3	0,0002	0,5363	−0,0001
4	0,0005	0,2483	0,0000
5	0,0004	0,4141	−0,0001
10	0,0004	0,4419	0,0000
15	0,0000	0,9791	−0,0002
30	−0,0012	0,1125	0,0003
60	−0,0032	0,0069*	0,0013

In the descriptive statistics, it became evident that there is a high correlation between the negative sentiment and the net sentiment score for both the BoW model and the RNTN. In the preliminary regression, the implication of this was that the results of the two different sentiment scores were similar in many ways regarding statistical significance. This inclination has not changed after the returns have been risk-adjusted. Similar conclusions on the model's explanatory power will, therefore, be drawn.

As it was the case with the two other sentiment scores, the adjusted R^2 is very low and almost 0 for the RNTN model. The BoW model does on the other hand still show signs of explanatory power, even though the adjusted R^2 is still very low after the returns has been risk-adjusted.
When the significance of the different coefficients is examined, the results are similar to the results of the negative sentiment scores. The negative sentiment scores of the BoW model were all significant at a 5% level, except for the coefficient at a window-size of 10 and 15 days, and the same is the case with the net sentiment score. The only significant coefficient for the RNTN, on the other hand, is at a window-size of 60 days.

6.4.4 Key Results

When the regressions are compared to the previous regressions, where the dependent variable was excess returns, then it is obvious that the significance of multiple coefficients has changed to become less significant and a part of these are now insignificant at a 5% significance level. An overview of the variables and whether they are significant is shown in Table 6.18. In addition, the adjusted R^2 is generally lower as well. This means that the sentiment scores were to a large degree explaining systematic risk instead of the return due to the sentiment of management. This is especially evident for the RNTN model, which, contrary to the BoW model, shows no significant results on the short window-sizes from 2 to 15 days. On the 60-day window-sizes, however, the variables are still significant and have the same mathematical sign - for both models. The sign remains contradictory of the intuitive interpretation – that a more positive (negative) sentiment equals an increase (decrease) in stock returns.

Looking at the results for the BoW model, it is evident that it is better at identifying trends in stock returns through its negative sentiment score variable than its positive score. This is evident by the larger number of significant variables for the negative sentiment score. This is consistent with Loughran and McDonald (2016, p. 1217) who describes that negative sentiment scores are less ambiguous since they are not influenced by framing issues.

Going forward, a choice will be made on which sentiment score to analyze. The negative and net sentiment score have overall outperformed the positive sentiment score in statistical significance, therefore, it makes sense that the choice stands between these two. There are, however, two disadvantages about the net sentiment score. Because of the high correlation between the negative and net sentiment score, as described in Section 6.2.4, it can be argued the net sentiment score is merely explaining the same phenomena as the negative sentiment score. In addition, the

positive sentiment score, which had low explanatory power, is included in the net sentiment score. Therefore, to avoid any influence by this variable, the negative sentiment score is viewed as the better choice.

To test whether other variables explain the variation in stock returns, the next steps will be to include various control variables in the regression. In order to do this, the negative sentiment score for both models will on a 2, 3, 4, 5, and 60-days windows-size be regressed, along with control variables, on the risk-adjusted returns. The reason for choosing these window-sizes is that this is where the negative sentiment score has shown the strongest results. In addition, the 60-day window-size is included to test whether the inclusion of additional variables explains the different mathematical sign of the coefficient of the variable.

Table 6.18: Coefficients of the different sentiment scores

Window-size	Positive Sentiment Score RNTN	Negative Sentiment Score RNTN	Net Sentiment Score RNTN	Positive Sentiment Score BoW	Negative Sentiment Score BoW	Net Sentiment Score BoW
2 Days	0,0001	−0,0001	0,0001	0,0024	−0,0059*	0,0052*
3 Days	0,0003	−0,0004	0,0002	0,0050	−0,0068*	0,0063*
4 Days	0,0007	−0,0007	0,0005	0,0033	−0,0053*	0,0048*
5 Days	0,0007	−0,0005	0,0004	0,0035	−0,0058*	0,0052*
10 Days	0,0005	−0,0007	0,0004	−0,0118	−0,0022	0,0008
15 Days	−0,0010	−0,0003	0,0000	−0,0299*	−0,0031	−0,0002
30 Days	−0,0041*	0,0011	−0,0012	−0,0409*	0,0088*	−0,0114*
60 Days	−0,0087*	0,0037*	−0,0032*	−0,0758*	0,0272*	−0,0304*

6.5 Risk-Adjusted Returns with Control Variables

In Section 5.7 it was described how there might be explanatory variables that could explain some of the variations in stock returns in addition to the sentiment scores. As described in Section 5.7 the control variables are the following. First of all, a proxy of the quality of earnings is used, which is the levels of net income and the change from the previous year in net income. The last control variable is a variable that describes whether the company is in the S&P 500 index or not.

These different control variables will be included in the regression simultaneously with the negative sentiment score. By having this approach, the impact of the sentiment in the 10-K report on stock returns is isolated. It will furthermore be examined if there is a difference between the BoW and RNTN models ability to explain movements in the risk-adjusted returns over the filing date. The results of these tests are shown and discussed below.

6.5.1 Bag of Words (BoW)

Table 6.19 shows the results of the BoW model.

Table 6.19: Regression results of BoW model including control variables

Window-size	2 days	3 days	4 days	5 days	60 days
Intercept	0,0015	0,0020	0,0003	0,0016	0,0011
Significance (P-value)	0,6356	0,4957	0,9167	0,6474	0,9070
Negative sentiment score	-0,0045	-0,0055	-0,0039	-0,0045	0,02122
Significance (P-value)	0,0580	0,0152*	0,1189	0,0027*	0,0032*
Levels of net income	0,0102	0,0177	0,0075	0,0054	-0,1230
Significance (P-value)	0,2902	0,0517	0,4599	0,6133	<0,0001*
Change in net income	0,0389	0,0339	0,0460	0,0383	-0,0987
Significance (P-value)	<0,0001*	0,0002*	<0,0001*	0,0003*	0,0006*
In S&P 500	0,0020	0,0030	0,0009	0,0021	-0,0162
Significance (P-value)	0,5052	0,3029	0,7844	0,5410	0,0803
Adjusted R^2	0,0055	0,0068	0,0060	0,0037	0,0121
F Statistic	<0,0001*	<0,0001*	<0,0001*	0,0002*	<0,0001*

In the table, the short-window sizes of 2 to 5 and the long window-size of 60 is examined. In the bottom of the table the adjusted R^2 and the F statistic is shown. First of all, the F statistic shows that it can be rejected that all the coefficients are equal to 0 simultaneously, which shows that the model is significant as a whole. When the adjusted R^2 is examined, it is shown that it increases compared to the regression that only included the risk-adjusted returns and the negative sentiment score. The inclusion of the control variables therefore increases the explanatory power of the regression.

The negative sentiment score variable is the variable of concern, however. Even though the control variables have been added to the regression, the negative sentiment score is still significant at most window sizes if a 5% significance level is used. The only exception to this is at the window-sizes of 2 and 4. At the 2-day windows-size the negative sentiment coefficient is significant at a 10% significance level and marginally close to being significant at a 10% level at the 4-day window-size. The signs of the negative sentiment coefficients have not changed by including the control variables since they are still negative at the short window-sizes and turn to positive at the 60-day window-size.

These results indicate that based on changes in the negative sentiment score for the BoW model, it is possible to explain risk-adjusted returns of short window sizes. It is on the other hand more questionable at long window-sizes such as 60 days, where the signal of the coefficient is opposite of the intuitive interpretation.

With regard to the control variables, it is clear that the change in net income has a positive and significant signal. The level of income has a positive coefficient as well but is mostly insignificant at the different window-sizes, however. Since the sentiment score does not take any quantitative data into account it makes sense that the change in net income is significant. The positive coefficient of the two income variables are, furthermore, as expected given a positive income and positive change in income will most likely be associated with positive stock returns.

The last variable is whether the companies are in the S&P 500 index at the filing date. This variable is not significant, which can be explained by the fact that there is still a lot of attention on the different companies if they have been in the index once, and it might therefore only have a minor effect if they are excluded from the S&P 500 index afterward.

In the end, it can be argued that the BoW model is able to extract incremental information from the 10-K reports by using the negative sentiment score. Which is the case even after controlling for the quality of earnings and whether the companies are in the S&P 500 index.

6.5.2 Recursive Neural Tensor Network (RNTN)

Table 6.20 shows the results of the negative sentiment score of the RNTN. The table is similar to that of the BoW model.

Table 6.20: Regression results of RNTN model including control variables

Window-size	2	3	4	5	60
Intercept	0,0013	0,0018	0,0002	0,0014	0,0016
Significance (P-value)	0,6675	0,5309	0,9419	0,6746	0,8604
Negative sentiment score	0,0001	-0,0002	-0,0005	-0,0003	0,0032
Significance (P-value)	0,9131	0,7418	0,3538	0,6214	0,0527
Levels of net income	0,0112	0,0091	0,0084	0,0065	-0,1282
Significance (P-value)	0,2428	0,0367*	0,4044	0,5435	<0,0001*
Change in net income	0,0403	0,0354	0,0468	0,0395	-0,1028
Significance (P-value)	<0,0001*	<0,0001*	<0,0001*	0,0002*	0,0003*
In S&P 500	0,0022	0,0031	0,0010	0,0022	-0,0167
Significance (P-value)	0,4825	0,2815	0,7602	0,5188	0,0705
Adjusted R^2	0,0048	0,0056	0,0056	0,0032	0,0110
F Statistic	<0,0001	<0,0001	<0,0001	0,0008	<0,0001

The F statistic of the regression shows that it can be rejected that all the coefficients are equal to 0 simultaneously, and the regression is therefore significant as a whole. The adjusted R^2 shows that the inclusion of the control variables increases the explanatory power of the regression, which was the case for the BoW regression as well.

When the P-values are examined, it is evident that negative sentiment scores of all window-sizes are insignificant, since it is rejected at a 5% significance level the negative sentiment score is different from 0. It is therefore also rejected at a 5% significance level that the negative sentiment score of the RNTN model gives any incremental information of the 10-K report. The coefficient that comes closest to being significant at a 5% level is at a window-size of 60 days. Given that this coefficient is unintuitive since it is positive it can be questioned how valid this variable is. The interpretation of the negative sentiment score for the RNTN model is, therefore, that it is not able to explain the variations in risk-adjusted returns over a filing date, which is the case at all window-sizes.

Based on the above regressions, it can be argued that the RNTN model is not able to extract incremental information from the 10-K report by using the negative sentiment score. When the negative sentiment score of the RNTN model was regressed as the only variable on the risk-adjusted returns, it was only the coefficient at the window-size of 60 days that were significant at 5% level. The ability of the RNTN model to explain the risk-adjusted returns over a filing date was therefore already questionable at this point. When the control variables were added to the regression, this lacking ability was exposed.

6.5.3 Comparison

When the significance of the negative sentiment score variable of the RNTN and BoW model was tested the same setup was used for both models. It is, therefore, possible to compare the explanatory power of the two coefficients.

First of all, the adjusted R^2 is larger at all window-sizes for the negative sentiment score of the BoW model. Furthermore, most of the coefficients of the BoW model were significant or close to be significant at a 5% level for the different window-sizes. Whereas none of the coefficients of the RNTN's negative sentiment score were significant at a 5% level. These results indicate that the BoW model is better at extracting incremental information from the 10-K report than the RNTN.

There can be different reasons for this result, but the major reason is probably that the BoW utilizes a dictionary that is made for a business context. This is not the case for the RNTN model. The implication is that even though the RNTN model is more advanced than the BoW since it examines complete sentence structures, the BoW model still performs better.

6.6 Test of Linear Regression Assumptions

To further assess the robustness of the results, it will be tested whether the assumptions of linear regression are fulfilled.

The tests will be performed on the regressions that included the negative sentiment score of the RNTN and BoW model, and the different control variables. The focus will be on the error term. First of all, it will be tested if the error term is normally distributed, has a mean of 0 and uniform variance, which is also called homoscedasticity. After

these tests have been made, it will be tested whether the error term is exposed to autocorrelation. In addition, it would be interesting to test whether the explanatory variables are exposed to multicollinearity, but this has already been tested previously in Section 6.2.4 where it was shown that this was not the case.

6.6.1 Mean Error

In the table below, it is shown that the mean error of the RNTN regression is zero for the window-sizes of 2 to 5 days, while there is a deviation from this at the window-size of 60 days. The BoW model shows similar results since the mean error only has minor deviations from 0, at the window-size of 2 to 5 days, while the deviation is larger and in the opposite direction at the window-size of 60 days.

Table 6.21: Mean error

Mean error by window-size	2	3	4	5	60
RNTN regression	0	0	0	0	0,0638
BoW regression	−0,0045	−0,0048	−0,0017	−0,0041	0,0312

6.6.2 Homoscedasticity

To test the assumption of homoscedasticity the following regression is used:

$$e_i^2 = a_0 + a_1 * \hat{y}_i$$

(Newbold, Carlson, & Thorne, 2013, p. 580)

$e_i^2 =$ The squared error term

$\hat{y}_i =$ The predicted values of the regression

This is called an auxiliary regression, where the R^2 coefficient of this regression is multiplied by the number of observation n, which gives the test statistics, which is shown in Table 6.22 and 6.23.

Table 6.22: Test of homoscedasticity RNTN model

RNTN	2	3	4	5	60
R^2 auxiliary regression	0,0011	0	0,0001	0,0002	0,0181
Observations	4772	4772	4772	4772	4772
Test statistic	5,0106	0	0,4772	0,9544	86,3732

Table 6.23: Test of homoscedasticity BoW model

BoW	2	3	4	5	60
R^2 auxiliary regression	0,0011	0	0,0001	0	0,0204
Observations	4772	4772	4772	4772	4772
Test statistic	5,1681	0	0,4772	0	97,3488

The test statistics in the above tables are compared to the Chi-Square Distribution. To be able to reject the hypothesis that the regressions have uniform variance over the predicted values, the test statistics have to be below 2,706 at a 10% significance level and below 5,024 at a 2,5% significance level. It is evident that not possible to reject the hypothesis for window-sizes of 3 to 5 days at a 10% significance level, and for the window-size of 2 days, the hypothesis is close to being rejected as well. The only case, where it is possible to reject the hypothesis is at a window-size of 60 days.

6.6.3 Autocorrelation

The last test that will be made is to see if the error terms in the regression model are correlated from one regression to the other. To test this auto-correlation, the Durbin-Watson test will be used. The Durbin-Watson test is shown in Table 6.24. The hypothesis is accepted if the calculated Durbin-Watson test statistic is larger than d_u and less than $4\text{-}d_u$, where d_u is defined as the cutoff points (Newbold, Carlson, & Thorne, 2013, pp. 584-585)
As seen in the table below, it can be rejected for both models at all window-sizes that no autocorrelation is present.

Table 6.24: Test for autocorrelation

Durbin-Watson Test	2	3	4	5	60
RNTN regression	1,9658	1,9636	1,9707	2,0279	1,9713
BoW regression	1,966	1,9651	1,9714	2,0288	1,9668
d_u	1,63	1,63	1,63	1,63	1,63
$4 - d_u$	2,37	2,37	2,37	2,37	2,37

6.6.4 Test of Normally Distributed Errors

To test if the errors are normally distributed the Jarque-Bera Test will be used. The test statistic of the Jarque-Bera Test is calculated in the following way:

$$Jarque - Bera = n * \left[\frac{(skewness)^2}{6} + \frac{(kurtosis - 3)^2}{24}\right]$$

(Newbold, Carlson, & Thorne, 2013, p. 611)

The results of the Jarque-Bera test are shown in the Table 6.25 and 6.26, where the hypothesis of normality is rejected if the test statistic exceeds the critical value. When the Jarque-Bera test statistic is compared to the critical value it becomes clear that hypothesis of normally distributed errors is rejected. The reason for the missing normality in the error terms usually arise from unusual data points, such as outliers, which were shown to be present in Section 6.2 regarding summary statistics earlier in this chapter. In the next section, it will be tested whether winsorizing of the most extreme values of the risk-adjusted returns will make the error term normally distributed and if this will change the results of the regressions considerably.

Table 6.25: Test for normally distributed errors RNTN model

RNTN	2	3	4	5	60
Skewness	24,1737	6,6714	7,5793	5,8418	2,8756
Kurtosis	1206,5300	227,2803	246,5734	174,8161	30,3540
Observations	4772	4772	4772	4772	4772
Jarque-Bera	288.007.091,04	10.001.652,75	11.796.393,81	5.869.719,23	148.776,69
Critical value (5%-point)	5,99	5,99	5,99	5,99	5,99

Table 6.26: Test for normally distributed errors BoW model

BoW	2	3	4	5	60
Skewness	24,1634	6,6647	7,5780	5,8390	2,8818
Kurtosis	1205,8483	227,0333	246,5400	174,7188	30,4067
Observations	4772	4772	4772	4772	4772
Jarque-Bera	287.680.919,17	9.979.635,23	11.793.158,87	5.863.073,03	149.350,51
Critical value (5%-point)	5,99	5,99	5,99	5,99	5,99

Based on these results, it can be concluded that the data suffers particularly from a non-normality of the error term, which might affect the robustness of the results. In Section 6.7 it will, therefore, be tested if this missing normality of the error terms is caused by outliers, and whether the assumption of normality is accepted when these outliers are treated. Finally, an assessment will be made of whether the winsorizing of the risk-adjusted stock returns will have any influence on the results.

6.7 Regression on Winsorized Risk-Adjusted Returns with Control Variables

In the previous section, it was described that the assumption of normally distributed errors was not satisfied. It is often the case with real data that unusual data points such as outliers are present, which might cause the assumption of normally distributed errors to be violated (Newbold, Carlson, & Thorne, 2013, p. 613)

Outliers were previously shown to be present in the risk-adjusted returns, and since these outliers are just extreme values and not caused by errors, then it is not possible to drop these observations. These outliers might have a large influence on the results though. Therefore, it will be examined if the errors can become normally distributed if these outliers are treated in another way. In addition, this test will reveal whether the influence of the outliers affects the coefficients in a way that makes them misleading, such as the coefficients of the 60-day window-size. To deal with these outliers a technique called winsorization will be used, where the reasoning behind this choice and how it is done, were described in Section 5.8.

After the risk-adjusted returns have been winsorized, they will be used as the dependent variable in the regressions. In this regression, the negative sentiment scores of the BoW and RNTN model, and the different control variables will be included. The results of these regressions are shown in Table 6.27, 6.28, 6.29, and 6.30.

6.7.1 Bag of Words (BoW)

When the risk-adjusted returns have been winsorized the result of the regression changes slightly in different ways. Of particular interest is the negative sentiment score and the adjusted R^2 since it will be examined if the explanatory power of the model has increased by winsorizing the risk-adjusted returns.

In the results of the regression, it is evident that the negative sentiment score at the window-size of 2 days has become significant at a 5% level. The negative sentiment score is also significant at the window-size of 60 days, however, at the window-sizes of 3 to 5 days the variable has become insignificant. The level of the coefficients for the negative sentiment score has furthermore changed, but the change is only minor, and the signal of the coefficients has not changed either.

If the adjusted R^2 is examined, then the general impression is that the adjusted R^2 has decreased compared to before the risk-adjusted returns were winsorized. It is furthermore shown in the Jarque-Bera test that even though the Jarque-Bera test statistic has decreased, it has not improved to a degree, where it can be accepted that the error terms are normally distributed.

Based on these results, it can be argued that the regression performed better before the returns were winsorized. The reason is that more coefficients are significant at 5% significance level, the level of the coefficients only change slightly, and there is no difference in the sign of the negative sentiment score between the regressions. Lastly, the adjusted R^2 is also higher in general, which adds to the argument that the explanatory power of the regressions on the risk-adjusted returns are higher.

Table 6.27: Regression results of BoW model on winsorized risk-adjusted returns

Window-size	2	3	4	5	60
Intercept	0,0012	0,0013	−0,0005	0	−0,0043
Significance (P-value)	0,3359	0,3854	0,8012	0,9882	0,5362
Negative sentiment score	−0,0022	−0,0013	−0,0014	−0,0019	0,0179
Significance (P-value)	0,0204*	0,2556	0,3751	0,2542	0,0009*
Levels of net income	0,0145	0,0177	0,0202	0,0190	−0,0405
Significance (P-value)	0,0002*	0,0001*	0,0011*	0,0046*	0,0637
Change in net income	0,0022	0,0035	0,0181	0,0146	-0,0535
Significance (P-value)	0,5568	0,4470	0,0032*	0,0275*	0,0133*
In S&P 500	0,0019	0,0021	0,0010	0,0016	-0,0120
Significance (P-value)	0,1179	0,2556	0,6317	0,4667	0,0847
Adjusted R^2	0,0050	0,0042	0,0059	0,0040	0,0058
F Statistic	<0,0001*	<0,0001*	<0,0001*	0,0001*	<0,0001*

Table 6.28: Test for normally distributed errors BoW model

BoW	2	3	4	5	60
Skewness	0,1135	0,1538	0,1535	0,1532	0,1734
Kurtosis	0,0972	0,0297	1,3858	1,1563	0,5084
Observations	4772	4772	4772	4772	4772
Jarque-Bera	1675,42	1754,25	518,09	675,88	1234,38
Critical value (5%-point)	5,99	5,99	5,99	5,99	5,99

6.7.2 Recursive Neural Tensor Network (RNTN)

The results of the regression on the negative sentiment score of the RNTN model also change slightly when the risk-adjusted returns have been winsorized. As it was the case with BoW model, there is a particular interest in the negative sentiment score and the adjusted R^2.

The results from the regression show that the negative sentiment score is still insignificant at a 5% level, which is the case at all window-sizes. It is furthermore shown that there are no major changes in the level or sign of the coefficients. Furthermore, the explanatory power of the model has decreased since the adjusted R^2 in general has decreased compared to before the risk-adjusted returns were winsorized.

In the Jarque-Bera test it is furthermore shown that even though the test statistic has decreased, it has not improved to a degree where it can be accepted that the error terms are normally distributed.

Based on these results, it can, therefore, be argued that both regressions perform poorly since none of the negative sentiment score coefficients are significant at a 5% level.

Table 6.29: Regression results of RNTN model on winsorized risk-adjusted returns

Window-size	2	3	4	5	60
Intercept	0,0011	0,0012	0,7906	0	−0,0039
Significance (P-value)	0,3626	0,3988	0,7906	0,9930	0,5816
Negative sentiment score	0	−0,0001	−0,0005	−0,0002	0,0024
Significance (P-value)	0,7211	0,6877	0,1431	0,5739	0,0515
Levels of net income	0,0150	0,0180	0,0205	0,0194	−0,0448
Significance (P-value)	<0,0001*	<0,0001*	0,0009*	0,0037*	0,0400*
Change in net income	0,0028	0,0038	0,0061	0,0150	−0,0573
Significance (P-value)	0,4534	0,4054	0,0030*	0,0231*	0,0080*
In S&P 500	0,0020	0,0021	0,0020	0,0016	−0,0124
Significance (P-value)	0,1078	0,1454	0,6174	0,4528	0,0734
Adjusted R^2	0,0039	0,0039	0,0062	0,0038	0,0043
F Statistic	0,0001*	0,0001*	<0,0001*	0,0002*	<0,0001*

Table 6.30: Test for normally distributed errors RNTN model

RNTN	2	3	4	5	60
Skewness	0,1085	0,1509	0,1535	0,1483	0,1862
Kurtosis	0,0934	0,0302	1,3856	1,1603	0,5179
Observations	4772	4772	4772	4772	4772
Jarque-Bera	1679,81	1753,66	518,22	672,95	1224,98
Critical value (5%-point)	5,99	5,99	5,99	5,99	5,99

6.8 Preliminary Conclusions

Two hypotheses were stated in order to structure the empirical examination of the thesis. These hypotheses have been tested by regressing the BoW's and the RNTN's sentiment scores of 10-K reports on stock returns. The results of these regressions, presented earlier in this chapter, allows for answers to these hypotheses.

H1: The Stanford CoreNLP software can be used to explain stock returns by analyzing the sentiment of 10-K reports.

The initial results of the analysis in this chapter show that while the adjusted R^2 is remarkably low there is a statistically significant relationship between the Stanford CoreNLP softwares' (RNTN) sentiment score and the stock returns. However, after testing the validity of the results through risk-adjusting the stock returns and adding control variables, the results of the RNTN showed no statistically significant relationship with changes in stock returns. This means the sentiment score merely explains the part of the variability of stock returns that is due to systematic risk and control variables, thus, providing no additional value. There are some issues with the robustness of data as mentioned in Section 6.6, which might affect the results of the analysis, however, the winsorizing process, that controls for this issue, yields the same results. These empirical findings suggest a rejection of the hypothesis that the RNTN can be used to predict stock returns by analyzing the sentiment of 10-K reports.

The second hypothesis in this Thesis is the following:

H2: The Stanford CoreNLP's sentiment analysis of 10-K reports is better at explaining stock returns than the Bag of Words approach using the Loughran & McDonald's (2011) financial dictionary to analyze the sentiment of 10-K reports.

Similar to the regressions of the RNTN's sentiment score, the initial results of the regressions on the BoW's sentiment score showed a statistically significant relationship between its output and the stock returns albeit with an extremely low adjusted R^2. Contrary to the RNTN's sentiment score, the BoW model's sentiment score, was marginally significant at the 2-day window-size and significant at the 3, 5, and 60-day window-size after risk-adjusting

the returns and adding control variables. As mentioned above, there are some issues with the robustness of data, nevertheless, the results are deemed to be valid. Overall, the conclusion is that the empirical findings suggest a rejection of the hypothesis that RNTN's sentiment analysis of 10-K reports is a better at explaining stock returns than the BoW approach using the Loughran & McDonald's (2011) financial dictionary to analyze the sentiment of 10-K reports.

The conclusions of the hypotheses enable an answer for the overall research question in this thesis, which was:

"To what degree can stock returns be explained by sentiment extracted from 10-K reports using the Stanford CoreNLP software and a Bag of Words approach."

Based on the conclusions of the hypotheses, the answer to the research question is that while stock returns cannot be explained by the RNTN, the BoW approach can. This is evident by the fact that the BoW's sentiment score showed a statistically significant relationship between its output and the stock returns albeit with an extremely low adjusted R^2. Even though the RNTN is a more sophisticated approach to sentiment analysis than the BoW approach, it seems that the choice of corpus has a larger say regarding the explanatory power of stock returns. This is evident by the fact that the RNTN misclassified neutral sentences characterized by technical language use which might consequently have led to a weak explanatory power of the model.

Chapter 7 – Discussion

This chapter will provide a discussion of the results and insights of this thesis, thus, giving a perspective of how to interpret the results, its limitations, and where and how to direct efforts in future research. First, a discussion of factors that may influence the results and the interpretation of them is presented. Lastly, the contribution of this thesis and suggestions for future research prospects is discussed.

7.1 Discussion of Results and Limitations

In this thesis, the RNTN and BoW approach were evaluated by their ability to explain stock returns over the filing date of 10-K reports. The results showed that the more sophisticated RNTN could not outperform the simple BoW model. This is likely due to the language domain mismatch in the sense that the model could not identify the neutral technical language of the 10-K reports and misclassified much of it as a negative sentiment.

When comparing the results with Feldman's findings (2009), it is clear that the results are quite similar. This is evident by the low R^2 and values of the coefficients that Feldman experienced as well (Feldman, Givindaraj, Livnat, & Segal, 2009, pp. 938, 946-947). Thus, this thesis confirms Feldman's (2009) findings that either the models used for quantitative analysis of qualitative information are still too crude or that the information in the narratives of financial reports do not have an impact on stock returns. Based on the knowledge about the tools applied in this thesis, which are a good representation of the current state of the art models for textual analysis, it seems, all things being equal, that the former might be the case. In addition, the fact that the simple BoW model achieved statistically significant results is an indication that there is still valuable information waiting to be retrieved in the qualitative information.

Reflecting on the analysis and its results it is clear that this thesis is largely affected by construct validity and inferred causality. Construct validity means there is an uncertainty regarding whether the proxies for phenomena really measure the underlying theoretical constructs they are intended to measure. When using these proxies to explain a dependent variable, it creates a problem with the inferred causality between the two variables. The problem is that since there is uncertainty about whether the proxies reflect the underlying phenomena there is uncertainty regarding the relationship between the phenomena and dependent variable. This is especially evident when using methods from the social sciences in the field of accounting research because many of the methods use proxies for phenomena that cannot be directly observed such as the level of emotion (DeFond, 2010, p. 404).

The implication of the problem with the construct validity in this thesis is that every test of the sentiment models' output is a joint test of the hypothesis that the output explains stock returns and the hypothesis that the output is a valid measure of sentiment. This means that while it is possible to conclude on the relationship between the output of the models and the stock returns, it presents difficulties in evaluating which models' output is "best" at expressing

the sentiment (DeFond, 2010, p. 404). For example, while the RNTN's output might be a valid proxy for the sentiment of management, this proxy might not correlate with stock returns as well as the BoW's proxy of sentiment. Without this in mind, one would be tempted to infer that the RNTN is not as good at extracting the sentiment of management as the BoW model, while in fact, this is not the conclusion of the analysis. While this is true, it is plausible that the RNTN could have performed better at extracting sentiment if it used a sentiment treebank made for financial text given its negative bias. This will be a discussion point later on.

Another factor to consider when interpreting the results is the analysis of the narratives in the whole 10-K report versus only the narratives in MD&A. If it was possible to parse the MD&A from the 10-K reports, the results might have been better, since the information from management about future and present performance is concentrated in the MD&A. This is, therefore, an area for future work since many researchers await the production of a range of accounting narratives in a structured, digitalized text format (Beattie, 2014, p. 128). This implementation might come soon since recent proposals from the SEC indicate that XBRL tagging in financial reporting may expand to other areas (EY, 2017, p. 11).

In addition, the information environment, which is defined as the aggregate of the entities that collect, process or disseminate information (and the extensiveness of the information itself), might have an influence on the results. Feldman (2009, p. 918) argues that the tone of the narratives in annual reports is less informative when the information environment around the firm is stronger. The reason for this is that a lot of the information in tone change has already been reflected in stock prices through the analysts' interpretations and interactions with management prior to the filing date (Feldman, Givindaraj, Livnat, & Segal, 2009, p. 945). In the S&P 500 index, which the analysis is based on in this thesis, the information environment is characterized by being very strong given the size and number of analysts following the firms. Because of this, the sentiment signal that has been sought in the analysis might already be incorporated in the stock price, which can explain some of the weak results.

7.2 Implications of the Results and Future work

By analyzing 10-K reports with a RNTN this thesis has taken a method and theory from computer science and used it in an accounting setting, thus applying theoretical pluralism. The results were mixed, however. Initially, the RNTN almost performed on par with the BoW model despite its large bias in negativity. However, when controlling for systematic risk and control variables the results revealed that the RNTN's sentiment score shows no statistically significant relationship with stock returns.

Given the RNTN's inability to identify much of the neutral technical language in the 10-K reports, it seems that it will benefit greatly from a sentiment treebank generated specifically for financial text. It is the opinion of the authors of this thesis that a sentiment treebank for finance will alleviate the RNTN's bias in negativity and provide a

powerful tailored model for sentiment analysis of financial text that compensates for the BoW model's shortcomings by taking grammar into consideration. Making such a corpus has additionally two advantages for the scientific community. Firstly, the incorporation of economic theory into the RNTN with words and sentences from financial settings will be a good example of further extending the theoretical pluralism advocated for in the field of accounting research (Beattie, 2014, p. 128). Second, a sentiment treebank for finance will provide a valuable contribution since accounting researchers await the creation of freely available communications corpora for finance of the type that exists in other fields of study (Beattie, 2014, p. 128).

An approach to creating a tailor-made financial sentiment treebank is replicating the Stanford Sentiment Treebank (Socher, et al., 2013). This would require a dataset like the one introduced by Pang and Lee (2005) except that instead of sentences from movie reviews, it should be based on sentences from the financial domain. These sentences could, for example, be from financial media and annual reports. Afterward, these sentences should be parsed into parse trees that show the syntactic structure of the sentences through the combination of phrases. Lastly, the phrases must be annotated by human judges. This is the most crucial step since it requires competent annotators and clear guidelines of how the phrases should be annotated. First, it is important that the annotators have enough knowledge about finance to be able to understand the technical language. Second, there are various characteristics of financial language that require certain conduct when annotating sentiment. An example could be that the sentiment classification should not be a proxy of whether the annotator believes that the sentence is related to a higher/lower stock return. While this would provide an interesting model, it is not exactly sentiment it would identify. It would also be misleading to use it in other studies than the prediction of stock returns. Another example is the case of performance-oriented statements. In general, performance-oriented statements can be negative, positive or neutral depending on their context and what they are compared against. However, given the sentiment model should be for the universal use and not just for a specific firm at a specific time, the annotator should assume that he/she has no contextual information whatsoever and classify such statements as neutral.

Following the methods of Socher et al. (2013) in building a sentiment treebank for a financial context and taking great care in the annotation process should provide researchers with a powerful tool that can to a higher degree analyze the complex linguistic phenomena and intricacies of the sentiment conveyed say, by managers in annual reports. In addition, this tool will help researchers in their work to understand the role and impact of accounting narratives in decision making.

While there are great benefits of developing such a model, there are increasingly more key stakeholders from investment banks, hedge funds and the SEC that are beginning to use and develop these kinds of sophisticated NLP algorithms to mine corporate communications for information relating to performance. Once preparers of financial reports realize this and they discover how the algorithms work, they are likely to engage in "reverse engineering" and

produce language that shows the features that will produce an outcome in their desire – be it a better cost of debt or stock returns. In response to this, the stakeholders would have to be ready to optimize their algorithms to reflect this change in language from preparers. The motivation behind this is whether it is worth the time and effort. In any case, this type of game-playing will likely erode the impact of the information in narratives and enhance the future challenge of better understanding the narrative choices of the human actors in external reporting and the consequences of these choices (Beattie, 2014, p. 127).

Chapter 8 - Conclusion

This chapter will provide a summary of the findings of this thesis. In addition, the main contributions of this thesis will be presented.

The overall aim of the thesis is to provide a comparison of the BoW model and the RNTN in their ability to explain stock returns over the 10-K filing date. This has been achieved by regressing their sentiment scores of the narratives in 10-K reports on stock returns. The initial results showed that while the adjusted R^2 was remarkably low, there was a statistically significant relationship between the models' sentiment scores and stock returns. The sentiment scores were on the short window-sizes over the filing date intuitive in their interpretation; an increase (decrease) in negativity leads to a decrease (increase) in stock returns. However, on the longer window-sizes, the relationship was the opposite. This relationship was consistent amongst both models. To assess the significance of the models the returns were adjusted for systematic risk using the Fama-French three factor model. When the returns were adjusted for systematic risk the significance of the coefficients worsened. This was especially the case on the short term window-sizes for the RNTN model, which showed no statistical significance. It was decided to further test on the negative sentiment score since it was the variable that showed the strongest statistical relationship with stock returns. This was done by regressing it together with various control variables. This regression revealed that while the RNTN's negative sentiment score showed no statistically significant relationship with stock returns, the BoW model's negative sentiment score was marginally significant at the 2-day window-size and significant on the 3, 5, and 60-day window-sizes.

The descriptive statistics revealed that there were many outliers in data. In addition, the tests of the linear regression assumptions revealed that some of these assumptions were not satisfied. The effect is that the coefficients and statistical significance of the regressions are not robust, which might affect the validity of the results. In order to test for the implications of the lacking robustness, a winsorization process of the data was initiated. This, however, did not yield the desired results since the coefficients and overall statistical significance did not change noticeably.

The overall conclusions are thus, that both models initially provided significant explanations of the stock returns over the filing date. However, after testing the validity of the relationships, it turned out that it is only the BoW model's sentiment score that provides significant results. As a result, it is evident that the two hypotheses, **H1** & **H2**, of this thesis, are both rejected. Based on the conclusions of the hypotheses, the answer to the research question is that while stock returns cannot be explained by the RNTN's sentiment analysis of the accounting narratives in 10-K reports, they can be explained through a BoW approach since this regression showed significant results.

8.1.1 The Contributions of this Thesis

The contribution of this thesis is that it has shown that the Stanford's Core NLP framework provides an option for analyzing the sentiment of financial reports on a sentence level. In addition, the Standford CoreNLP framework allows for easy replication of the results and overall usage for future studies. However, as an explainer of stock returns, the results are questionable and certainly not as good as conventional models rooted in economic theory such as the BoW model.

Following the methods of Socher et al. (2013) in building a sentiment treebank for a financial context should provide researchers with a powerful tool that to a higher degree can analyze the intricacies of the sentiment conveyed, say by managers in annual reports. In addition, this tool will help researchers in their work in understanding the role and impact of accounting narratives in decision making. However, the development of sophisticated NLP algorithms will likely result in game-playing between key-stakeholders and preparers that might erode the impact of the information in narratives.

Appendix 1 – Mail from Bill McDonald

Answer from Bill Mcdonald

I do not have parsed data with the financials.

Parsing MDA's accurately is, in my opinion, virtually impossible and yet everybody claims to do it. If firms all followed the standard rules in terms of the form structure it would not be difficult, but they do not. In addition, many times the MD&A is put into an exhibit which is introduced in section 7, sometimes with enough introductory comments to make it difficult to determine where the MD&A is. I am very skeptical of research that leans heavily on this parse, but I know it is frequently done. Good luck.

Bill

Bill McDonald

Professor of Finance | Thomas A. and James J. Bruder Chair in Administrative Leadership
335 Mendoza College of Business | University of Notre Dame | Notre Dame, IN 46556
P: 574-631-5137 | E: mcdonald@nd.edu | W: http://www.nd.edu/~mcdonald

Question for Bill McDonald

Hi again Bill,

Thank you for adding me to the list.

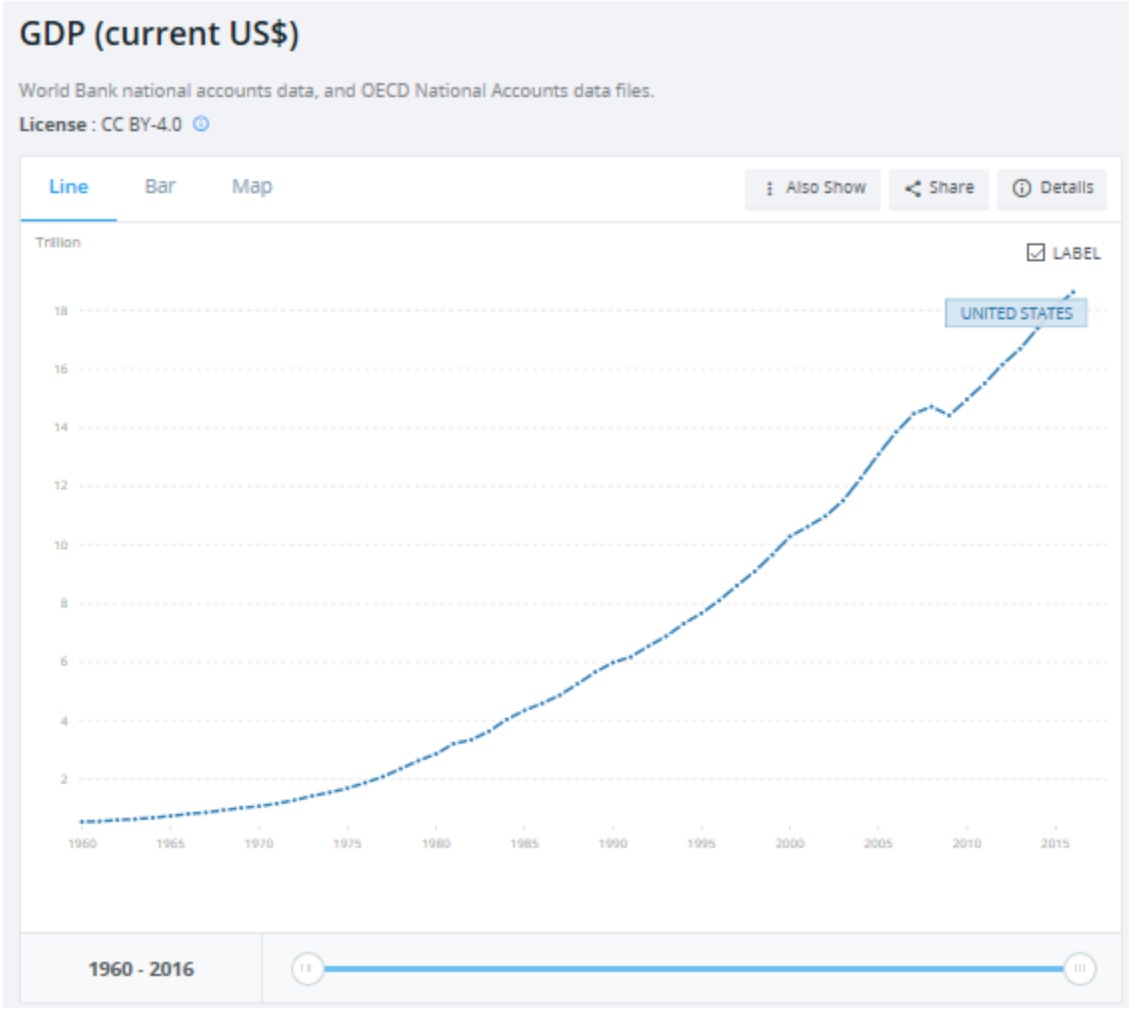
I'm writing again, as we're having difficulties with the data collection.

Of course we have access to the XBRL filings available on EDGAR, but one of the main items that we want to examine is the MD&A - and since this item is not a mandatory field in XBRL, it is not included in the files. We have downloaded the zip-file from your online database (very helpful!) and had a look at the data. As it contains all the text from quarterly reports but no financial data, we are faced with the problem of joining the data we have from Edgar with you data, and then we furthermore need to find an efficient way of parsing out the MD&As (which can prove to be difficult due to the possible inconsistency in companies' way of including it). Thus, before we begin this work, I thought I might write you and ask whether you have a data set that combines all the content from the quarterly reports (that is, financial variables and text)?

I know it's a long shot, and if not, I would still like to express my gratitude for you making the textual data available online - that's a great help to researchers.

Appendix 2 – GDP USA

Figure A.2: Development in GDP USA



(The World Bank, 2018)

Appendix 3 – Stop Words

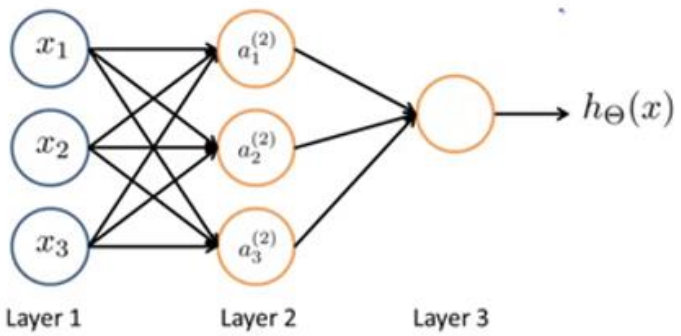
ME MY MYSELF WE OUR OURS OURSELVES YOU YOUR YOURS YOURSELF YOURSELVES HE HIM HIS HIMSELF SHE HER HERS HERSELF IT ITS ITSELF THEY THEM THEIR THEIRS THEMSELVES WHAT WHICH WHO WHOM THIS THAT THESE THOSE AM IS ARE WAS WERE BE BEEN BEING HAVE HAS HAD HAVING DO DOES DID DOING AN THE AND BUT IF OR BECAUSE AS UNTIL WHILE OF AT BY FOR WITH ABOUT BETWEEN INTO THROUGH DURING BEFORE AFTER ABOVE BELOW TO FROM UP DOWN IN OUT ON OFF OVER UNDER AGAIN FURTHER THEN ONCE HERE THERE WHEN WHERE WHY HOW ALL ANY BOTH EACH FEW MORE MOST OTHER SOME SUCH NO NOR NOT ONLY OWN SAME SO THAN TOO VERY CAN JUST SHOULD NOW

Appendix 4 – Neural Networks and their Training

A4.1 Model Representation and Notation

A neural network with 3 inputs, 3 hidden units, and one output layer, would be illustrated like such:

Figure A4.1: A Simple Neural Network



Source: (Ng, Coursera, 2018a)

Where layer 1 is the input layer, layer 2 is the hidden layer, and layer 3 is the output layer. The arrows act as "weights" or the neural networks parameters and are denoted θ. Vectorization is a big deal in neural networks, since it dramatically reduces the computation time for training the model. Therefore, Θ^j is denoted as the matrix of weights controlling function mapping from layer j to layer j+1, and X as the matrix of inputs.

Θ in this example is thus: $\begin{bmatrix}\theta_0\\ \theta_1\\ \theta_2\\ \theta_3\end{bmatrix}$, and $X = \begin{bmatrix}x_0\\ x_1\\ x_2\\ x_3\end{bmatrix}$, where θ_0 and x_0 are bias units. The bias units allow for "movement" the sigmoid curve, such that it fits the prediction with the data better. It works similar to the constant *b* of a linear function: *y = ax + b*. Without *b* the line always goes through (0,0) and might give a poorer fit.

In this example there is only one hidden layer, but a neural network can have any number of hidden layers. The neurons (units) in the layers are generally denoted as a_i^j which means the "activation" of unit i in layer j. There are a few activation functions, one of the most common is the Sigmoid (logistic) activation function, which is denoted as: $g(z) = \frac{1}{1+e^{-z}}$. An explanation of what z denotes will be given below.

The following show how the activation functions a_i^j are written out:

$$a_1^2 = g(\Theta_{10}^1 x_0 + \Theta_{11}^1 x_1 + \Theta_{12}^1 x_2 + \Theta_{13}^1 x_3)$$

$a_2^2 = g(\Theta_{20}^2 x_0 + \Theta_{21}^2 x_1 + \Theta_{22}^2 x_2 + \Theta_{23}^2 x_3)$

$a_3^2 = g(\Theta_{30}^3 x_0 + \Theta_{31}^3 x_1 + \Theta_{32}^3 x_2 + \Theta_{33}^3 x_3)$

And the hypothesis output is given by:

$h_\Theta(x) = a_1^3 = g(\Theta_{10}^2 a_0^2 + \Theta_{11}^2 a_1^2 + \Theta_{12}^2 a_2^2 + \Theta_{13}^2 a_3^2)$

This is saying that the computation of the activation nodes is done by using a 3x4 matrix of parameters. Each row of the parameters is applied to the inputs to obtain the value for one activation node. The hypothesis output is the logistic function applied to the sum of the values of the activation nodes, which have been multiplied by yet another parameter matrix Θ^2 containing the weights for the second layer of nodes (Ng, Coursera, 2018a).

The dimensions of these matrices are determined as follows: if a network has s_j units in layer j, s_{j+1} units in layer j+1, then Θ^j will be of dimension $s_{j+1} * (s_j + 1)$. The +1 comes from the addition of in Θ^j of the "bias nodes", x_0 and Θ_0^j.

To vectorize the above function, a new variable z_k^j is defined that covers the parameters inside the g function. Thus, if all the parameters are replaced with z this will be the result:

$a_1^2 = g(z_1^2)$

$a_2^2 = g(z_2^2)$

$a_3^2 = g(z_3^2)$

Setting $x = a^1$, z can be rewritten z as: $z^j = \Theta^{j-1} a^{j-1}$. Now, a vector of the activation units for layer j can be derived: $a^j = g(z^j)$.

The bias unit to layer j is added after a^j is computed. This will be element a_0^j and will be equal to 1. To compute the final hypothesis, $h_\Theta(x)$, another z vector will first be computed: $z^{j+1} = \Theta^j a^j$, which is done by multiplying the next Theta matrix (after Θ^{j-1}) with the activation values of all the activation nodes. This last Theta matrix Θ^j will have only one row which is multiplied by one column a^j (the one unit in layer 3). Thus, the result is a single number (z^j). The final result (the hypothesis output) is derived with: $h_\Theta(x) = a^{j+1} = g(z^{j+1})$. The intuition of the sigmoid hypothesis output is: the estimated probability that y=1 on input x. Adding all these intermediate layers in neural networks allows for more elegant production of interesting and more complex non-linear hypotheses on a large amount of data. (Ng, Coursera, 2018a)

A4.1.1 Neural Network Backpropagation

The cost function, $J(\Theta)$ is denoted as:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{K}\left[y_k^{(i)}\log\left(\left(h_\theta(x^{(i)})\right)_k\right) + \left(1-y_k^{(i)}\right)\log\left(1-\left(h_\theta(x^{(i)})\right)_k\right)\right] + \frac{\lambda}{2m}\sum_{l=1}^{L-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_l+1}\left(\theta_{j,i}^{(l)}\right)^2$$

Where, L = the total number of layers in the network.

$s_l =$ number of units (not counting the bias unit) in layer 1.

K = number of output units

The double sum adds up the logistic regression costs calculated for each cell in the output layer, and the triple sum adds up the squares of all the individual Θ's in the entire network.

The goal is to compute $\min_{\Theta} J(\Theta)$, or in other words, to minimize the cost function J using an optimal set of parameters in theta. To do this, a computation of the partial derivative of $J(\Theta)$: $\frac{\partial}{\partial\Theta_{i,j}^{l}}J(\Theta)$ is done using an algorithm such as gradient descent that iterates until the minimum of $\frac{\partial}{\partial\Theta_{i,j}^{l}}J(\Theta)$ is found (Ng, Coursera, 2018a).

A4.1.2 RNTN Backpropagation

This section from Socher et al (2013) describes in detail the training of the RNTN model. Each node in the model has a softmax classifier that is trained through the vector representation to predict a target vector t. The target vector t is assumed to be a one-hot vector with a number of classes (sentiments) as length (C) - a 1 at the correct label and all other values to be 0. The predicted vector y will be the probabilities of the classes and is a product of the classifiers.

The goal is to maximize the probability of the correct prediction. This is done by minimizing the cross-entropy error between the predicted distribution $y^i \in R^{C*1}$ at node i and the target vector $t^i \in R^{C*1}$ and that node.

Following the cross-entropy method, the cost function with the RNTN parameters $\theta = (V, W, W_s, L)$ for a sentence is: $E(\theta) = \sum_i \sum_j t_j^i \log y_j^i + \lambda\left|\left|\theta\right|\right|^2$

Where V is the multiple bilinear form that transforms the word and phrase vectors into vectors that reflect the compositional value of the two. W is the main parameter for learning, L is the word embedding matrix, and W_s is the classification matrix.

From here, $x^i \in R^{d*1}$ is defined as the word/phrase vector at node i. Each node backpropagates its error through to the recursively used weights V and W. The softmax error vector at node i is defined as: $\delta^{i,s} \in R^{d*1}$. The error vector is calculated as: $\delta^{i,s} = (W_s^T(y^i - t^i)) \otimes f'(x^i)$, where $\otimes$ is the Hadamard product between the two vectors and f' is the element-wise derivative of f.

The full derivative for V and W is the sum of the derivatives at each of the nodes, which are computed in a top-down fashion from the top node through the tree and into the leaf nodes.

The following is an illustration of the calculation of the error vector: the complete incoming error messages for a node i is defined as $\delta^{i,com}$. The top node, here p_2, only received errors from its softmax, therefore, $\delta^{P_2,com} = \delta^{P_2,s}$. For the derivative of each slice $k = 1, \dots, d$ is given by:

$$\frac{\partial E^{p_2}}{\partial V^{[k]}} = \delta_k^{p_2,com} \begin{bmatrix} a \\ p_1 \end{bmatrix} \begin{bmatrix} a \\ p_1 \end{bmatrix}^T,$$

Where $\delta_k^{p_2,com}$ is the k'th element of this vector. This is used for computing the error message for the two children of p_2:

$\delta^{p_2,down} = (W^T \delta^{p_2,com} + S) \otimes f'\left(\begin{bmatrix} a \\ p_1 \end{bmatrix}\right)$, where S is defined as:

$$S = \sum_{k=1}^{d} \delta_k^{p_2,com} (V^{[k]} + (V^{[k]})^T \begin{bmatrix} a \\ p_1 \end{bmatrix}$$

The children of p_2 will then each take half of this vector and add their own softmax error message for the complete δ.

In particular: $\delta^{p_1,complete} = \delta^{p_1,s} + \delta^{p_2,down}[d+1:2d]$,

Where $\delta^{p_2,down}[d+1:2d]$ indicates that p_1 is the right child of p_2 and hence takes the 2nd half of the error, for the final word vector derivative for a, it will be $\delta^{p_2,down}[1:d]$.

The full derivative for slice $V^{[k]}$ for this trigram tree then is the sum at each node:

$$\frac{\partial E}{\partial V^{[k]}} = \frac{\partial E^{p_2}}{\partial V^{[k]}} = \delta_k^{p_1,com} \begin{bmatrix} b \\ c \end{bmatrix} \begin{bmatrix} b \\ c \end{bmatrix}^T$$

and similarly for W. For this nonconvex optimization the AdaGrad algoritm (Duchi, Hazan, & Singer, 2011) has been used (Socher, et al., 2013).

Appendix 5 – Development in Average Sentiment Scores

Table A5.1: Development in Average Sentiment Scores

	2009	2010	2011	2012	2013	2014	2015	2016	2017
Negative Sentiment Score BoW	0,2243	0,0091	0,0056	0,0038	0,0126	0,0126	0,0265	0,0350	0,0189
Positive Sentiment Score BoW	−0,0163	−0,0141	0,0135	0,0148	−0,0038	0,0157	0,0080	0,0015	0,0015
Net Sentiment Score BoW	−0,2406	−0,0232	−0,0079	0,011	−0,0164	0,0031	−0,0185	−0,0336	−0,0173
Negative Sentiment Score RNTN	0,4506	−0,2439	−0,1012	−0,0303	−0,0884	0,0337	−0,0468	0,1742	0,0032
Positive Sentiment Score RNTN	−0,1467	0,1801	0,1729	0,0798	0,097	0,0276	0,0041	−0,0605	−0,0213

Appendix 6 – Second Parsing

Based on manual search in several texts we found points where the data quality had potential to be improved. We therefore picked out 50 random 10-K from which we identified elements that added noise to the text. From the 50 reports we found the following: Most of the reports had page numberings which in a digital form of the reports appear as number on a single line. We remove the newline and the page numbers. Some reports used roman page numbers instead. We also remove the newlines and the roman numbering. Some of the reports had markers around the page numbers. We also remove these examples. We removed all tabulating characters. Tabulates are often used in table of contents of the reports and to structure other tables throughout the reports. Loughran and McDonald have removed most of the tables, but some of the tabulating characters remained. Removing the tabulate characters have no effect on the analysis as the later word tokenizing also include tabulates as stop character. We further removed all appearances of the word 'PART' in upper case. Companies use 'PART' as an indicator for the beginning of a new part. We removed any numeric or roman numbers that followed 'PART' We finally removed all html tagging used by Loughran and McDonald to indicate the header of a report and any exhibits (Jønsson & Jakobsen, 2018).

Bibliography

Adelberg, A. H. (1979). Narrative disclosures contained in financial reports: means of communication or manipulation? *Accounting and Business Research*, 179-189.

Ahmad, K., Han, J. G., Hutson, E., Kearneyd, C., & Liu, S. (2016). Media-expressed negative tone and firm-level stock returns. *Journal of Corporate Finance*, 152–172.

Audi, R., Loughran, T., & McDonald, B. (2015). Trust, but Verify: MD&A Language and the Role of Trust in Corporate Culture. *Journal of Business Ethics*, 551-561.

Baroni, M., & Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 673–721.

Beale, A. (2018). *Version 6 of the 12dicts word lists*. Retrieved March 22, 2018, from http://wordlist.aspell.net/12dicts-readme/#classic

Beattie, V. (2014). Accounting narratives and the narrative turn in accounting research: Issues, theory, methodology, methods and a research framework. *The British Accounting Review* , 111-134.

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A Neural Probabilistic Language Model. *Journal of Machine Learning Research* , 1137–1155.

Bernard, V. L., & Thomas, J. K. (1989). Post-Earnings-Announcement Drift: Delayed Price Response or Risk Premium? *Journal of Accounting Research*, 1-36.

Bodie, Z., Kane, A., & Marcus, A. J. (2014). *Investments*. Maidenhead: McGraw-Hill Education.

Brau, J. C., Cicon, J., & McQueen, G. (2016). Soft Strategic Information and IPO Underpricing. *Journal of Behavioral Finance*, 1-17.

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

Cardie, C., Choi, Y., & Breck, E. (2007). Identifying expressions of opinion in context. *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, (pp. 2683–2688).

Clarkson, P. M., Kao, J. L., & Richardson, G. D. (1999). Evidence That Management Discussion and Analysis (MD&A) is a Part of a Firm's Overall Disclosure Package. *Contemporary Accounting Research*, 111-134.

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: deep neural networks with multitask learning. *Proceedings of the 25th international conference on Machine learning*, (pp. 160-167).

Core, J. E. (2001). A review of the empirical disclosure literature: discussion. *Journal of Accounting and Economics*, 441-456.

CRSP. (2018a). *About CRSP*. Retrieved April 20, 2018, from http://www.crsp.com/about-crsp

CRSP. (2018b). *CRSP Daily Stock*. Retrieved April 20, 2018, from https://wrds-web.wharton.upenn.edu/wrds/query_forms/variable_documentation.cfm?vendorCode=CRSP&libraryCode=crspa&fileCode=dsf&id=ret

CRSP. (2018c). *Data Definitions - P*. Retrieved April 20, 2018, from www.crsp.com/products/documentation/data-definitions-p

Cutler, D. M., Poterba, J. M., & Summers, L. H. (1989). What moves stock prices? *Journal of Portfolio Management*, 4–12.

Damodaran, A. (2018a). *Damodaran Online*. Retrieved May 3, 2018, from Estimating Risk free Rates: http://people.stern.nyu.edu/adamodar/pdfiles/papers/riskfree.pdf

Damodaran, A. (2018b). *Damodaran Online*. Retrieved May 3, 2018, from Estimating Risk Parameters: http://people.stern.nyu.edu/adamodar/pdfiles/papers/beta.pdf

Danske Bank Group. (2018). *Annual Report 2017*.

Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web. *Management Science*, 1375-1388.

Davidson, J. (2008). Repitition, rhetoric, reporting and the 'dot.com era: words, pictures, intangibles. *Accounting, Auditing and Accountability Journal*, 791-826.

Davis, A. K., Piger, J. M., & Sedor, L. M. (2006). Beyond the numbers: An analysis of optimistic and pessimistic language in earnings press releases. *Contemporary Accounting Research*, 1-39.

DeFond, M. L. (2010). Earnings quality research: Advances, challenges and future research. *Journal of Accounting and Economics*, 402–409.

Diction. (2015, July 1). *Diction - The Text-Analysis Program*. Retrieved March 22, 2018, from Diction Overview: https://www.dictionsoftware.com/diction-overview/

Dow Global. (2017, September 1). *DowDuPont Merger Successfully Completed*. Retrieved April 12, 2018, from https://www.dow.com/en-us/news/press-releases/dowdupont-merger-successfully-completed

Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 2121-2159.

Erk, K., & Padó, S. (2008). A structured vector space model for word meaning in context. *EMNLP 08 Proceedings of the Conference on Empirical Methods in Natural Language Processing*, (pp. 897-906).

EY. (2017). *2017 SEC annual reports - Form 10-K*. Ernst & Young LLP.

Fama, E. F., & French, K. R. (1992). The Cross-Section of Expected Stock Returns. *The Journal of Finance*, 427-465.

Federal Reserve Bank of ST. Louis. (2018). The Financial Crisis - Full Timeline. St. Louis. Retrieved April 11, 2018

Feldman, R., Givindaraj, S., Livnat, J., & Segal, B. (2009, August 20). Managements's tone change, post earnings announcement drift and accruals. *Review of Accounting Studies*, 915-953. Retrieved April 24, 2018

Ferris, S. P., Hao, G. Q., & Liao, M.-Y. S. (2013). The Effect of Issuer Conservatism on IPO Pricing and Performance. *Review of Finance*, 993–1027.

Francis , J., & Schipper , K. (1999). Have financial statements lost their relevance? *Journal of Accounting Research*, 319-352.

French, K. R. (2018, March 29). *Data Library.* Retrieved May 3, 2018, from Current Research Returns: http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

Ghosh, A., & Moon, D. (2005). Auditor Tenure and Perceptions of Audit Quality. *The Accounting Review*, 585-612.

Ghosh, D., & Vogt, A. (2012). Outliers: An Evaluation of Methodologies.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning.*

Grefenstette, E., Dino, G., Zhang, Y.-Z., Sadrzadeh, M., & Baroni, M. (2013). Multi-step regression learning for compositional distributional semantics. *Proceeding of the 10th International Conference on Computational Semantics* (pp. 131-142). Potsdam: Association for Computational Linguistics .

Henry, E. (2008). Are investors influenced by how earnings press releases are written? *Journal of Business Communication*, 363-407.

Henry, E., & Leone, A. J. (2014). Measuring the Tone of Accounting and Financial Narrative. In E. Henry, & A. J. Leone, *Communication and Language Analysis in the Corporate World* (pp. 36-47). Hershey: IGI Global.

Huang, E. H., Socher, R., Manning, C. D., & Ng, A. Y. (2012). Improving Word Representations via Global Context and Multiple Word Prototypes. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* (pp. 873-882). Association for Computational Linguistics.

Irsoy, O., & Cardie, C. (2014). Opinion mining with deep recurrent neural networks. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, (pp. 720–728).

Jegadeesh, N., & Wub, D. (2013). Word power: A new approach for content analysis. *Journal of Financial Economics*, 712-729.

Jones, M. J., & Clatworthy, M. (2003). Financial reporting of good news and bad news: evidence from accounting narratives. *Accouting and Business Research*, 171-185.

Jønsson, K. R., & Jakobsen, J. B. (2018, May 15). Predicting Stock Performance Using 10-K Filings.

Kalchbrenner, N., Grefenstett, E., & Blunsom, P. (2014). A convolutional neural network for modelling sentences. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, (pp. 655–665).

Kearney, C., & Liu, S. (2014). Textual Sentiment in Finance: A Survey of Methods and Models. *International Review of Financial Analysis*, 171-185.

Klein, D., & Manning, C. D. (2003). Accurate Unlexicalized Parsing. *ACL 03 Proceedings of the 41st Annual Meeting on Association for Computational Linguistics.*

Kothari, S. P., Li, X., & Short, J. E. (2009). The Effect of Disclosures by Management, Analysts, and Business Press on Cost of Capital, Return Volatility, and Analyst Forecasts: A Study Using Content Analysis. *The Accounting Review* , 1639-1670.

Krippendorff, K. (2004). Content analysis: An introduction to its methodology (2nd ed.). Beverly Hills, CA: Sage Publications.

Kvålseth, T. O. (1985). Cautionary Note About R^2. *The American Statistician*, 279-285.

Lapata, M., & Padó, S. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, 161–199.

Lev, B., & Ohlson, J. A. (1982). Market-Based Empirical Research in Accounting: A Review, Interpretation, and Extension. *Journal of Accounting Research*, 249-322.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics*, 221–247.

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

Liu, P., Joty, S., & Meng, H. (2015). Fine-grained Opinion Mining with Recurrent Neural Networks and Word Embeddings. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, (pp. 1433–1443).

Loughran, T., & McDonald, B. (2011). When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, 35-65.

Loughran, T., & McDonald, B. (2012). IPO first-day returns,offer price revisions, volatility, and form S-1 language. *Journal of Financial Economics*, 307-326.

Loughran, T., & McDonald, B. (2015). The Use of Word Lists in Textual Analysis. *The Journal of Behavioral Finance*, 1-11.

Loughran, T., & McDonald, B. (2016, September 4). Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research*, 1187-1230.

Loughran, T., & McDonald, B. (2017). The Use of EDGAR Filings by Investors. *Journal of Behavioral Finance*, 231-248.

Luo, Y., & Zhou, L. (2017). Managerial ability, tone of earnings announcements, and market reaction. *Asian Review of Accounting*, 454-471.

Lynn, S. (2018). *Data science, Startups, Analytics and Data visualisation*. Retrieved April 2018, from https://www.shanelynn.ie/get-busy-with-word-embeddings-introduction/

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, (pp. 55-60).

McDonald, B. (2018a). *10-X_C_Zipped*. Retrieved April 12, 2018, from https://drive.google.com/drive/folders/0B4niqV00F3mscFpoRFVzVE9aM0E

McDonald, B. (2018b). *Software Repository for Accounting and Finance*. Retrieved March 22, 2018, from https://sraf.nd.edu/textual-analysis/resources/#Master%20Dictionary

McDonald, B. (2018c). *Stage One 10-X Parse Data*. Retrieved April 7, 2018, from https://sraf.nd.edu/data/stage-one-10-x-parse-data/#_ftn1

McDonald, B., & Loughran, T. (2011). Barron's Red Flags: Do They Actually Work? *Journal of Behavioral Finance*, 90-97.

McKinsey&Company. (2017). *Artificial Intelligence The Next Digital Frontier?* McKinsey&Company.

McLeish, D. L. (2005). *Monte Carlo Simulation and Finance.* Hoboken: John Wiley & Sons.

Merkl-Davies, D. M., Brennan, N. M., & McLeay, S. J. (2011). Impression management and retrospective sense-making in corporate narratives: A social psychology perspective. *Accounting, Auditing & Accountability Journal*, 315-344.

Mikolov, T., Corrado, G., Chen, K., & Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR*, 1–12.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *Proceedings of the 26th International Conference on Neural Information Processing Systems*, (pp. 3111–3119).

Mitchell, J., & Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science*, 1388–1429.

Mitchell, T. M. (1997). *Machine Learning.* McGraw-Hill.

Mitra, S., & Hossain, M. (2009). Value-relevance of pension transition adjustments and other comprehensive income components in the adoption year of SFAS No. 158. *Review of Quantitative Finanance and Accounting*, 279-301.

Morrison, A. (2015, October 1). *Quora*. Retrieved April 2018, from https://www.quora.com/What-do-people-who-work-in-or-with-big-data-think-about-limits-of-quantitative-data

Newbold, P., Carlson, W. L., & Thorne, B. M. (2013). *Statistics for Business and Economics.* Harlow: Pearson Education Limited.

Ng, A. (2018a). *Coursera.* Retrieved January 3, 2018a, from https://www.coursera.org/learn/machine-learning/home/info

Ng, A. (2018b). *Syllabus and Course Schedule.* Retrieved Januar 03, 2018, from http://cs229.stanford.edu/syllabus.html

Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *ACL 05 Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* , (pp. 115-124).

Pedersen, L. H. (2015). *Efficiently Inefficient: how smart money invests and market prices are determined.* Princeton University Press.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: global vectors for word representation. *Proceedings of the 2014 Empirical Methods in Natural Language Processing*, (pp. 1532–1543).

Petersen, C. V., & Plenborg, T. (2012). *Financial Statement Analysis.* Harlow: Pearson Education Limited.

Roll, R. (1988). R-Squared. *The Journal of Finance*, 541–566.

S&P Dow Jones Indices. (2018, April 12). *S&P 500*. Retrieved April 12, 2018, from https://eu.spindices.com/indices/equity/sp-500

S&P Global. (2018, March). *S&P Dow Jones Indices.* Retrieved 11 April, 2018, from S&P U.S. Indices Methodology: https://us.spindices.com/documents/methodologies/methodology-sp-us-indices.pdf

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, 210-229.

Sarkar, D. (2016). *Text Analytics with Python*. New York: Springer Science+Business Media.

Schmidhuber, J., & Hochreiter, S. (1997). Long short-term memory. *Neural computation*, 1735–1780 .

Shiller, R. (1981). Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends? *American Economic Review*, 421-436.

Socher, R., & Manning, C. (2018, January 24). *Natural Language Processing with Deep Learning*. Retrieved February 15, 2018, from http://web.stanford.edu/class/cs224n/archive/WWW_1617/lectures/cs224n-2017-lecture14-TreeRNNs.pdf

Socher, R., Lin, C.-Y., Ng, A. Y., & Manning, C. D. (2011). Parsing natural scenes and natural language with recursive neural networks. *Proceedings of the 28th International Conference on International Conference on Machine Learning*, (pp. 129-136).

Socher, R., Manning, C. D., & Huval, B. (2012). Semantic compositionality through recursive matrix-vector spaces. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, (pp. 1201-1211).

Socher, R., Manning, C. D., & Ng, A. Y. (2010). Learning Continuous Phrase Representations and Syntactic Parsing with Recursive Neural Networks. *Proceedings of the Deep Learning and Unsupervised Feature Learning Workshop of NIPS 2010*, (pp. 1-9).

Socher, R., Pennington, J., Huang, E. H., Ng, A. Y., & Manning, C. D. (2011). Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, (pp. 151-161).

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. *Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, (pp. 1-12).

Soper, F. J., & Dolphin Jr., R. (1964). Readability and corporate annual reports. *The Accounting Review*, 358-362.

Stanford. (2009, July 4). *Stemming and lemmatization*. Retrieved April 1, 2018, from https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

Stanford. (2018a). *Stanford CoreNLP – Natural language software*. Retrieved April 1, 2018, from https://stanfordnlp.github.io/CoreNLP/index.html

Stanford. (2018b). *Stanford Log-linear Part-Of-Speech Tagger*. Retrieved April 1, 2018, from https://nlp.stanford.edu/software/tagger.html

Stanford. (2018c). *Stanford Tokenizer*. Retrieved April 1, 2018, from https://nlp.stanford.edu/software/tokenizer.html

Stanford. (2018d). *The Stanford NLP Group*. Retrieved April 1, 2018, from https://nlp.stanford.edu/

Stanford. (2018e). *The Stanford Parser: A statistical parser*. Retrieved April 1, 2018, from https://nlp.stanford.edu/software/lex-parser.html: https://nlp.stanford.edu/software/lex-parser.html

Sun, S., Luo, C., & Chen, J. (2017). A review of natural language processing techniques for opinion mining systems. *Information Fusion*, 10–25.

Tai, K. S., Socher, R., & Manning, C. D. (2015). Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, (pp. 1556–1566).

Tetlock, P. C. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal Of Finance*, 1139-1168.

The World Bank. (2018). GDP (current US$). Retrieved April 11, 2018, from https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?end=2016&locations=US&name_desc=true&start=1960

Toutanova, K., Klein, D., Manning, C., & Singer, Y. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology* , (pp. 252-259).

Trim, C. (2013, January 23). *IBM developerWorks*. Retrieved March 16, 2018, from https://www.ibm.com/developerworks/community/blogs/nlp/entry/tokenization?lang=en

Tsai, F.-T., Lu, H.-M., & Hung, M.-W. (2016). The impact of news articles and corporate disclosure on credit risk valuation. *Journal of Banking & Finance* , 100–116.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 141–188.

U.S. Securities and Exchange Commission. (2003, December 19). *Interpretation: Commission Guidance Regarding Management's Discussion and Analysis of Financial Condition and Results of Operations*. Retrieved April 17, 2018, from https://www.sec.gov/rules/interp/33-8350.htm

U.S. Securities and Exchange Commission. (2009a, June 26). *Annual report pursuant to section 13 or 15(d) of the securities exchange act of 1934*. Retrieved April 4, 2018, from Form 10-K: https://www.sec.gov/about/forms/form10-k.pdf

U.S. Securities and Exchange Commission. (2009b, June 26). *Form 10-K*. Retrieved April 4, 2018, from https://www.sec.gov/fast-answers/answers-form10khtm.html

U.S. Securities and Exchange Commission. (2011, July 1). *How to Read a 10-K*. Retrieved April 4, 2018, from https://www.sec.gov/fast-answers/answersreada10khtm.html

U.S. Securities and Exchange Commission. (2014, Octiober 15). *Annual Report*. Retrieved April 4, 2018, from https://www.sec.gov/fast-answers/answers-annrephtm.html

U.S. Securities and Exchange Commission. (2017, August 2). *EDGAR Company Filings | CIK Lookup*. Retrieved April 12, 2018, from https://www.sec.gov/edgar/searchedgar/cik.htm

WRDS. (2018a, April 12). *Compustat Daily Updates - Index Constituents*. Retrieved April 12, 2018, from https://wrds-web.wharton.upenn.edu/wrds/ds/compd/index/constituents.cfm?navId=83

WRDS. (2018b). *WRDS Overview of CRSP/COMPUSTAT Merged (CCM)*. Retrieved April 20, 2018, from https://wrds-web.wharton.upenn.edu/wrds/support/Data/_001Manuals%20and%20Overviews/_002CRSP/ccm-overview.cfm

Yekini, L. S., Wisniewski, T. P., & Millo, Y. (2016). Market reaction to the positiveness of annual report narratives. *The British Accounting Review*, 415-430.

Yessenalina, A., & Cardie, C. (2011). Compositional matrix-space models for sentiment analysis. *Conference on Empirical Methods in Natural Language Processing* (pp. 172-182). Edinburgh: Association for Computational Linguistics.

Zanzotto, F. M., Fallucchi, F., Korkontzelos, I., & Manandhar, S. (2010). Estimating linear models for compositional distributional semantics. *Proceedings of the 23rd International Conference on Computational Linguistics* , (pp. 1263-1271).