Erasmus University Rotterdam

Erasmus School of Economics

Master Thesis Econometrics and Management Science

Predicting a firm's future state by utilizing a textual analysis approach containing a choice of words model and sentiment analysis.

Author:

Diederik Portheine (582581)

Supervisor Erasmus university: dr. MH Akyuz

Supervisor Accenture: P van Vreeswijk

Second assessor: dr. F Frasincar

Date: July 31, 2022

Abstract

The amount of daily created data is growing at an increasing speed. This research focuses on financial data of companies, more specific annual reports of companies in the manufacturing and IT industry of the United States. Analysts and investors are looking for means to analyse this vast amount of data. Because this amount is increasing, the demand to automate this process has never been higher. Due to the increase in computational power new approaches that include machine learning and deep learning are made possible. One of these new approaches is textual analysis, which can be described as extracting information from text by utilising an algorithm. However, textual analysis is a new field for financial documents and much research still needs to be done. This paper is an addition to the growing number of studies focused on financial documents. In addition to standing literature, a choice of words model is combined with a sentiment analysis to answer the question of whether new and valuable information can be extracted from the texts in annual reports for making predictions on a company's health or financial situation. In addition to this contribution, the approach is constructed such that the resulting model is highly interpretable. After all, "black-box" machine learning models are often forsaken in finance because these models are inherent too complicated to understand. The used annual reports are dated between 1995 and 2020 and in 2018 the dataset is divided into a training and test set. A lasso regression is trained to select these words that can predict the future state of a company. A subsequent factor analysis will reduce these dimensions and result in highly interpretable factors. The found factors are combined with the result of the sentiment analysis and this resulting model is used to predict the total revenues of the test set. It is found that the addition of sentiment increases the variation in the model and has a slight positive effect on the accuracy for short term predictions. The downside however is that it drastically decreases the accuracy of long term predictions.

Furthermore, the proposed model is compared with standard financial prediction methods, fitting a polynomial on historical data. It is found that for the manufacturing industry the proposed model scores similar on the short term predictions but has better accuracy on the long term predictions. However, for the IT industry the model performs slightly worse for long term predictions. This is accounted to the differences in the two industries.

Nevertheless, it can be concluded that the texts in annual reports contain quantifiable information which can improve current financial prediction methods. The model still can be further improved as it neglects the fact that observations are companies in a certain fiscal year. Implementing a panel data model would utilize the fact that these observations are linked in time and company which could greatly increase performance. This is due to the fact that different companies in the same industry can be adjusted for accordingly.

This research has contributed to promoting the use of textual analysis in finance. It can be seen that with further improvements this model can be used to improve financial predictions. The highly interpretable factors create new insights and the usage of annual reports which are widely accessible makes implementing these prediction methods simple and therefore desirable.

Contents

1 Introduction

Nowadays, technology has become a deeply embedded part of our lives. Nearly all activities in modern life are made possible due to technological advancements (Panchiwala and Shah 2020). Humanity is still improving its ways and technological advances are growing exponentially. One of the consequences of technology being a big part of our everyday lives is the amount of daily data created and used. According to the National Security Agency of the United States of America almost 2000 petabytes on average are handled daily over the internet (Jaseena, David, et al. 2014). With the increase in computational power and the exponentially growing data, various techniques have been developed to aid the analysis of this vast amount of information. These techniques range from classification, summarising and improving the ease of access and management of the data (Talaviya et al. 2020). Currently, a lot of time and interest is devoted to machine learning and deep learning, which is suited to handle these vasts amounts of data. A large part of the available data is in the form of text. An old field in research that has recently been rediscovered due to the increase in computational power is textual analysis, often referred to as text mining or data mining. Text analysis is the process of deriving valuable information and patterns from text. In accounting and finance, applying textual analysis could save analysts and investors a lot of time. Textual analysis can be applied to all sorts of financial sources like firm specific news, conference calls, Securities and Exchange Commission (SEC) filings and 10-Ks or annual reports. However, implementing textual analysis in financial documents is still an emerging area where much research is yet to be done. Financial data is found to be substantially in the form of text and most of the time unstructured. Multiple pieces of research concluded that the implementation of textual analysis has numerous applications in the finance industry, such as various kinds of predictions, customer relationship management, and cybersecurity issues, among others (Gupta et al. 2020). Many novel methods have been proposed for analyzing financial results in recent years. Artificial intelligence has made it possible to analyze and predict financial outcomes based on historical data (Culotta, Ravi, and Cutler 2016). These methods are often criticized due to complexity of the used models. In finance, prediction accuracy is desired but not at the expense of the interpretability of the used method. Financial data contains a significant amount of latent information. If the latent information were to be extracted manually from a vast corpus of data, it might take years. Advancements in text mining have made it possible to examine financial textual data efficiently.

Investigation of a company often starts with reading its financial disclosures. One of these disclosures is the annual report which contains a lot of different items and financial statements. From a financial point of view, making accurate predictions about a company and its trajectories is vital and investors often choose to merely focus on the financial statements for these predictions. However, this raises the question if this focus is justified or is there information present in the texts of annual reports which is neglected. This research will help answer the question if the texts found in annual reports contain any new and usable information that can be extracted by means of textual analysis for predicting a companies future state. To answer this question one needs to determine what information is useful to extract from the text. From literature it is clear that the different types of information that can be extracted are abundant. For

example, one could extract tone, sentiment, specific words, length, complexity and more. This research will focus on the sentiment and the choice of words found in the Management Discussion and Analysis (MD&A) filings in annual reports. The extracted information will be used to predict future company metrics to see if real usable information is present. These predictions will be compared with predictions based on the financial statements. To extract the information in the text an approach based on textual analysis is applied to the MD&A filings of companies in the manufacturing and IT industry. The used approach needs to be interpretable and understandable or it will not be implemented for financial predictions. In this research the proposed approach consists of a lasso regression with subsequent factor analysis to find which word frequencies can determine these latent drivers in the text. This combination satisfies the requirement because the found factors can easily be understood and the essence of the lasso regression is quite intuitive unlike many "black-box" machine learning methods.

In addition to standing literature, sentiment analysis results will be added to improve the predictions. The sentiment analysis is based on a Bidirectional-Encoder-Representations-from-Transformers (BERT) algorithm trained to understand financial disclosures. BERT is a powerful algorithm that often is trained for sentiment analysis but it can be trained for much more. For example, BERT has been trained to read financial headlines on news articles and Twitter feeds (Jaggi et al. 2021). However, applying BERT to annual reports as done in this research is something that has not often been done. Combined with the factor analysis, much information is extracted and this approach could potentially save analysts and investors time by automating the process of analyzing annual reports.

In summary, the sentiment and choice of words in MD&A filings of companies operating in the manufacturing and IT industry is extracted by means of textual analysis. The used approach consists of two parts. Part one is about extracting the choice of words which is done by means of a lasso regression with a subsequent factor analysis to reduce the amount of chosen words and find the latent drivers to predict future total revenues and EBITDA, see Section 2. The resulting factors are interpretable and contain more insights than only the predictions. The second part of the approach extracts the sentiment of the MD&A filings. The results of this sentiment analysis are combined with the found factors to predict total revenues and EBITDA. These predictions are compared with common prediction methods in finance and it is found that MD&A filings contain new and useful information that when extracted can substitute common prediction methods.

This thesis consists of multiple sections. A brief introduction to the used corporate finance theory will be given in the following section. In the subsequent section recent literature and research on textual and sentiment analysis will be presented and discussed. This is done in a more general sense and also for specific research on financial documents. In the methodology section, the applied methods will be elaborated upon and explained. This will be followed by a Data section where the used data will be discussed. After that, the approach will be applied to the data and the results will be presented and compared with other existing prediction methods. To conclude this research, the results will be reflected upon and discussed and possibilities for future research will be proposed.

2 Background in Corporate Finance

In this section the company metrics, total revenue and EBITDA, that are predicted is briefly explained to create a feeling of why these metric are important in the world of finance. In addition, the commonly used prediction method will be presented as it will be used as a comparison with the used approach of this research.

2.1 Total Revenue

The first metric, is in its essence a fairly straightforward metric. Revenue indicates how much a company receives in money from the products or services they provide. The formal definition for total revenue is the following:

> *"The combination of all incoming sources of money that the company has earned through the selling of goods or services. Total revenue is calculated as an average sales price per good or unit multiplied by the number of goods or units sold. Total revenue does not consider expenses directly related to the cost of making the good, delivering the product or service, or general operating expenses like salaries, taxes, or utilities."* (Mankiw 2021)

Even though total revenue is not necessarily indicative of overall profits, it is still vital to business owners. Total revenue gives insights into how much money a company is making and therefore provides an indication of a company's place in the market and its current state. It is known that total revenues are in general "sticky" as it is called in finance. This means that the total revenues do not deviate much from past total revenues and changes are often gradual. Multiple methods are applicable to forecast total revenues. However, as financial data does not necessarily need to be published it is hard to predict the revenue based on its fundamentals like total sales and sale prices. Analysts often try and extrapolate based on past total revenues by means of a trendline or moving average line. This paper uses extrapolation based on fitting a high dimensional function through historical data points to predict future total revenues. As total revenues are generally sticky this often will result in good predictions. The extrapolated data will be compared with the predicted data to see if the proposed approach is promising or is a good addition to a model based on past total revenues. This will be elaborated upon in the results in Section 6.

2.2 EBITDA

The second metric, earnings before interest, taxes, depreciation, and amortization, or in short EBITDA is a measure of a company's overall financial performance. It is one of the key metrics investors use when determining the health of a company. EBITDA, however, can be misleading because it does not account for the costs and only reflects on the economical potential of a company. It does not reflect the cost of capital investments like property, plants, and equipment. The formal definition is given below.

> *"A company's earnings before interest, taxes, depreciation, and amortization (commonly abbreviated EBITDA) is a measure of a company's profitability of the operating business only, thus before any effects of indebtedness, state-mandated payments, and costs required to maintain its asset base. It is derived by subtracting from revenues all costs of the operating business (e.g. wages, costs of raw materials, services ...) but not decline in asset value, cost of borrowing, lease expenses, and obligations to governments."* (Grant and Parker 2002)

Simply put, EBITDA is a measure of profitability. EBITDA is one of the metrics that is widely used and company's have no legal requirement to disclose. However, it can often be worked out and reported using the information found in a company's financial statements. EBITDA is a vital metric for investors because it indicates the financial prospects of a company by indicating a company's overall performance with its current assets. The correct way to estimate EBITDA is by estimating all different components of EBITDA. However, in practice this is often quite troublesome as data is not available to the public. Therefore analysts resort to, as is often done in finance, extrapolating from historical data to estimate future EBITDA scores. As done with total revenues, the extrapolated data will be compared with the predicted data to see if the proposed approach is promising or is a good addition to a model based on past EBITDA scores. This will be elaborated upon in the results in Section 6.

3 Literature review

A general process of textual analysis can be decomposed into three steps: harvesting text, cleaning and parsing the text, and analyzing the text. In terms of harvesting financial text different sources can be used. Common approaches are to collect data from financial disclosures, websites or social media. Researchers often choose to collect data from a financial database, such as Thomson Reuter News Database, newspaper database or EDGAR which is the database of the SEC (Rogers, Skinner, and Zechman 2017). Textual data is often unstructured and hard to understand for a computer. Cleaning and parsing the data is often referred to as pre-processing the text to a specific data format. This results in cleaned text which is easier to process for a computer. Only after harvesting and preprocessing can any type of textual analysis be done. In the field of textual analysis, researchers can gain different insights based on other techniques. Nowadays with machine learning, researchers can train software to classify text or recognize patterns. The methods often used for these problems are Naive Bayes, multiple linear regression, support vector machine and neural networks (Gupta et al. 2020). In Figure 1, the most popular approaches employed by accounting and finance researchers are schematically shown but more methods exist.

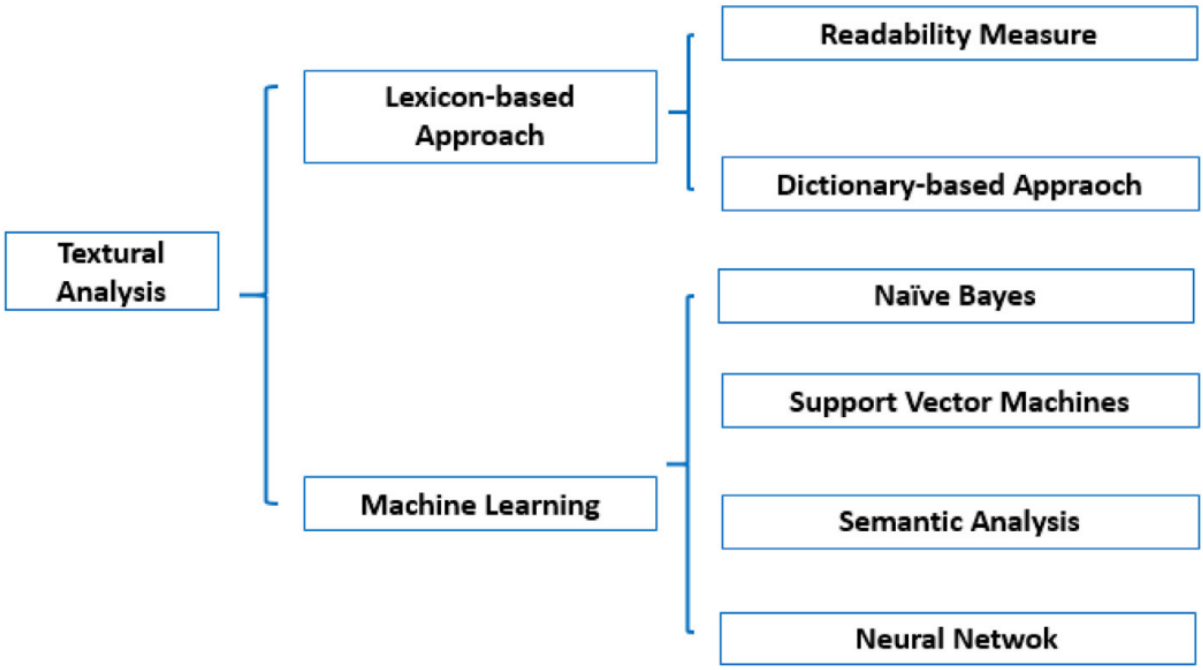

Figure 1: General methods for textual analysis. (Guo, Shi, and Tu 2016)

In the following subsection, Section 3.1, will the different applications of textual analysis be presented. The focus will be made on literature about financial documents as this research is applied to financial documents. After the introduction to all that is possible with textual analysis the subsequent section, Section 3.2, will be presenting past research about sentiment analysis. In this research the used model for sentiment analysis is BERT which is transformer based and therefore this section will be focused on literature containing similar models. In the last part of the literature review, Section 3.3, the literature about word based textual analysis will be discussed. The separation in this section is made because although the sentiment analysis and word-based analysis are both forms of textual analysis they are inherently very different as will become clear in this section.

3.1 Textual analysis in finance

The field of textual analysis is quite old. Still, it has been given a new life due to the increase in computation power, which also made implementing machine learning techniques possible. One of the first textual analyses was applied to the letters of the president and it was found that confidence in presidential statements was correlated with higher stock returns (Sanger and McConnell 1986). Nowadays, textual analysis in finance is often applied to find multiple metrics to gain different insights than the ones present in the figures of financial documents. An example that is often done is quantifying the tone of corporate disclosures. Tone or sentiment has a tangible impact on the decisions of stakeholders such as investors, analysts, and auditors. First, numerous studies find that the tone of financial disclosures (e.g., earnings announcements, Forms 10-Q (quarterly report) and 10-K (annual report) increases short-term stock returns and reduces stock return volatility (Kothari, Shu, and Wysocki 2009), even though this market reaction reverses over the long horizon (Huang 2014). In addition, it is found that around the time the forms 10-Q and 10-K are filed, market reactions are more positive when the tone in these disclosures is more positive. This analysis is also done with the transcripts of conference calls and similar results were found (Price et al. 2012). Another highly investigated source of information for textual analysis is social media. One of the first use cases of textual analysis in finance on social media is done by analyzing messages from Yahoo stock forums and comparing this with short-term stock returns (Das and Chen 2007). The social media of finance is often said to be Twitter, which is very applicable for textual analysis due to its limitation in message length. Twitter has already proven itself useful in different cases. For instance, Twitter messages were used to gauge earthquake intensity (Burks, Miller, and Zadeh 2014), lessening negative news from CEOs (Elliott, Grant, and Hodge 2018) and more. For finance it is no different as it is found that Twitter sentiment is highly correlated with turnover, mini flash-crash count and the number of trades at an intra-day level (Agrawal et al. 2018). Another time-consuming process in finance is fraud detection where textual analysis is very applicable. The research focused on the Management Discussion and Analysis (MD&A) section of a 10-K form, whether word usage differs between Accounting and Auditing Enforcement Releases (AAER) firms and industry-age-size matched non-AAER firms. They created a fraud score variable that is fitted using in-sample filings from 1997 to 2001 and used the 2002 to 2010 time period as an out-of-sample test. Firms that exhibit similarity with the abnormal vocabulary associated with AAER companies have a significantly higher probability of ex-post

accounting misstatements. In other words, specific vocabulary choices can help predict accounting fraud (Hoberg and Lewis 2017). Lastly, a different implementation of textual analysis is classifying corporate sustainability reports (CSR). This has become crucial from the financial reporting perspective. This is because the manual analysis is time-consuming and a new automated model was investigated (Shahi, Issac, and Modapothala 2014). The result was a text-mining approach with a more intelligent scoring of CSR reports. After preprocessing the dataset, four classification algorithms were implemented, namely Naive Bayes, random subspace, decision table, and neural networks. Naive Bayes with the Correlation-based Feature Selection filter was chosen as the preferred model. Based on this model, software was designed for CSR report scoring that lets the user input a CSR report to get its score as an automated output. The software was tested and had a high overall effectiveness.

3.2 Sentiment analysis

Sentiment analysis has a wide variety of different use cases. To give a brief introduction, a recent research that discusses different text-mining algorithms widely used in accounting and finance is presented (Guo, Shi, and Tu 2016). They merged the Thomson Reuters News Archive database and the News Analytics database. The former provides original news, and the latter includes sentiment scores ranging from -1 to 1 with negative, neutral, and positive scores. To balance the dataset, 3000 news articles were randomly selected for training and 500 for testing. Three algorithms, namely naive Bayes, SVM, and neural network, were run on the dataset. With the neural network having the highest accuracy, it was concluded that it could be used for text mining-based finance studies. A different upcoming field in sentiment analysis is natural language processing (NLP). Big companies are investing heavily in the opportunities of NLP. Being able to teach a computer to read and understand sentences is something that is highly desired. One of these companies involved in the development is Google and they created BERT (Devlin et al. 2018). With the introduction of BERT, researchers without the funding and resources can tap into the resources Google has. BERT is categorized as a language representation model. Unlike recent language representation models, BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right contexts in all layers. In other words, BERT can learn a language from unlabeled text which is in abundance. This initial learning step is done by Google. Google has trained BERT on a lot of textual data and as a result, the pre-trained BERT model can be fine-tuned with little additional training by often training one additional output layer. This makes it possible to create state-of-the-art models tailored to the task at hand. Nowadays, BERT is widely applied to all sorts of research. One research was conducted on the public opinion in social media by using the pre-trained BERT but learning it to focus on negative sentiment to classify same entities (Zhao et al. 2021). Because BERT can read faster and substantially more than a human and in finance, speed is often the key to success. BERT is widely adopted. A researcher found that training BERT to classify news articles can provide valuable information to predict stock market movements which also proved to be profitable (Sousa et al. 2019). Because BERT already knows how to read, it is not difficult to teach him how to read very specific types of documents. One finds that databases and structured data are often sparse or very expensive, this statement holds for financial databases. With BERT it is possible to

train the algorithm to read financial documents without large amounts of data. In this research BERT will be tailored to read financial documents as is done in this research (Yang, Uy, and Huang 2020). The steps taken will be elaborated upon in Section 4.1 .

3.3 Word based analysis

Word based analysis is focused on the choice of words in texts. This is not to be confused with the dictionary based approach which is often referred to as bag-of-words. The critical difference is that word based analysis does not use pre-labeled dictionaries but relies on the coherence between words. Different properties can be extracted from investigating the choice of words. One recent study has researched the relation between text in financial disclosures and a company's performance. The result was that companies with good financial performance and bad performing companies often use different types of words. It was found that the kind of words that appear predominantly in the business text is correlated with a company's sales performance. (Lee et al. 2018). Another research focused on the footnotes found in financial disclosures because these often contain useful information. Classifying these footnotes can be a cumbersome job. However, it was found that these unstructured footnotes could be classified using word based analysis and therefore training a model to classify these footnotes could save analysts lots of time (Heidari and Felden 2015).

4 Methodology

The methodology section will consist of multiple steps where the objective is to predict the future state of a company. In this section, the followed steps will briefly be introduced, whereas in the subsequent sections these introduced steps will be elaborated upon. The different phases and steps are schematically shown in Figure 2. The first part of the research will consist of setting up the sentiment analysis. An open source machine learning framework does the sentiment analysis for NLP called BERT. To correctly implement BERT it needs to be trained and fine-tuned to be used for the sentimental analysis of financial documents. The final BERT algorithm is referred to as FinBERT. As mentioned in Section 3, the process of textual analysis can be decomposed into three steps, namely, harvesting-, pre-processing-, and analyzing the text. Data harvesting will consist of gathering MD&A sections of 10-K filings using the EDGAR database. These filings are pre-processed to make them readable by algorithms and are combined with their corresponding Compustat data, see data Section 5. The next step consists of conducting the textual analysis. First, the sentiment analysis will be conducted, resulting in three values that indicate the sentiment of the text. Afterward, as mentioned earlier the future state of a company will be predicted. A good proxy for this are the total revenue and EBITDA, these two metrics will be referred to as company metrics. To create an insightful and interpretable analysis a lasso regression will be implemented to find words that contain information to help predict this future state. After the lasso regression is finished a factor analysis will be applied to reduce the dimensions of the results. These factors in combination with the results of the sentiment analysis will be used to determine if these texts contain predictive power. To conclude from the extracted information, a linear regression will be implemented to find out if the factors and sentiment contain additional information besides variables stated in the financial statements which are typically used in financial predictions.

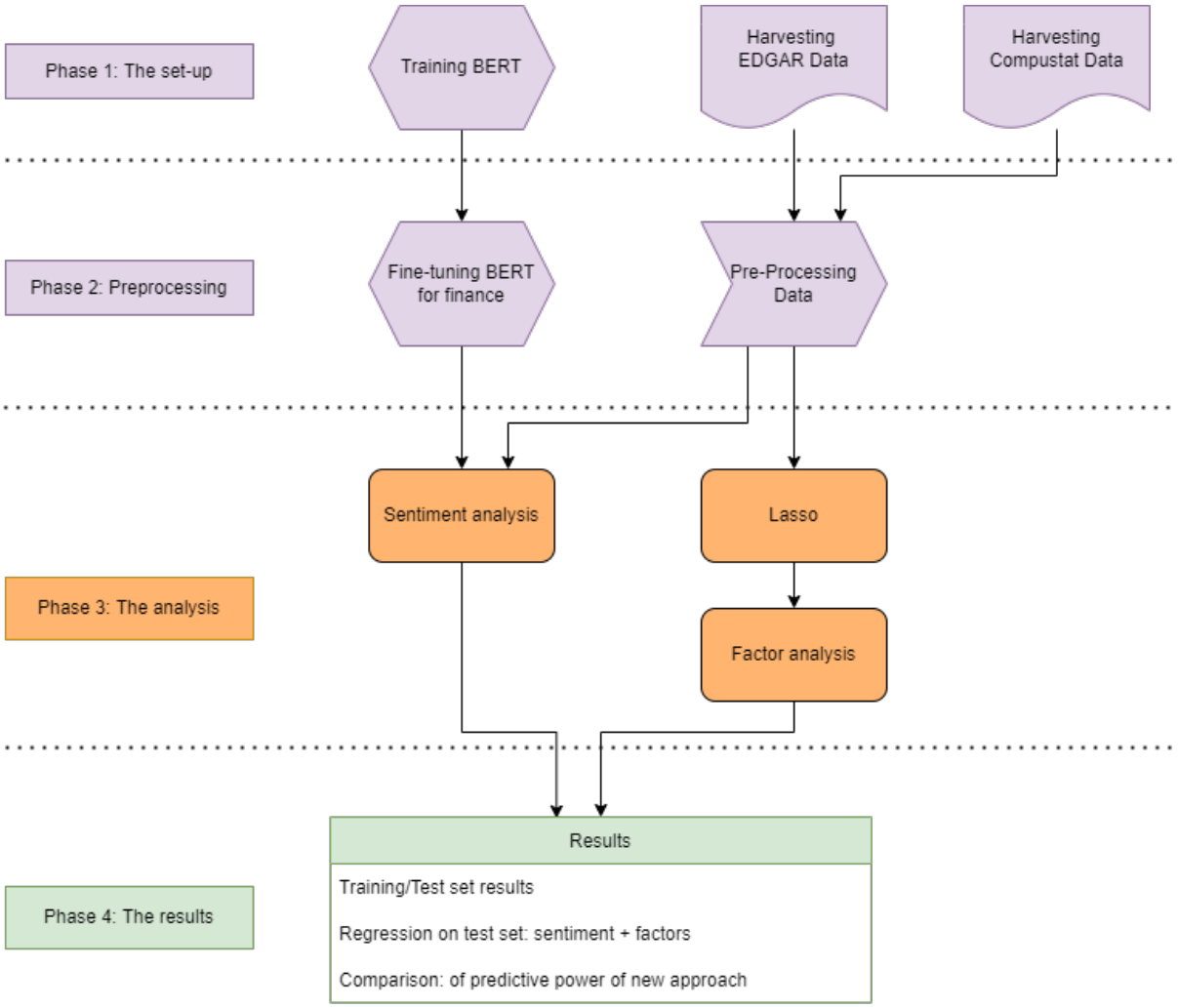

Figure 2: Flowchart containing the steps of this research.

As mentioned earlier, the used approach needs to be understood by investors with no prior computational linguistic knowledge. The combination of well established statistical procedures in textual analysis, Lasso followed by a factor analysis facilitates this. The lasso regression has the property that as selection operator it is very suited for the word-based model. This is because lots of words need to be removed because they lack in predictive power for the estimated company metrics. In addition, the broad operation of the lasso regression is easy to grasp which makes it understandable. The factor analysis is needed to further reduce the dimensions. Lasso selects from all the words those words that contain any predictive power and the factor analysis finds combinations in these words to make the result more interpretable. Rich interpretation of the resulting factors can help investors better understand what the resulting output captures. To further improve the accuracy of the predictions, the sentiment analysis results are combined with the factors. The sentiment model is BERT and the underlying structure of this model is very complex. However, it is made in such a way that it can easily be trained and the results understood. In addition, sentiment analysis in general is understandable and easy to grasp. In broad sense the approach can be simplified down to the following, the Lasso regression as a least absolute shrinkage

and selection operator selects words that predict company metrics. The subsequent factor analysis finds combinations of words which have broadly the same effect on these predictions and result in factors with rich interpretation due to the labeling property based on the underlying words. In addition, the sentiment of the MD&A section is added to further improve prediction accuracy. This combination of found factors and sentiment is used to predict future company metrics.

4.1 Training FinBERT

Before harvesting the data, the first thing to do is train BERT to be applied in sentiment analysis on financial documents. A sophisticated sentiment analysis is highly recommended because naive approaches such as bag-of-words would not cut it. This is mainly because essential context information, of which financial documents are full of, is discarded in these approaches. BERT is one of these advanced techniques which can crack the financial context and understand and define what is negative and what is positive from a financial point of view. BERT is relatively new and its fundamental technology relies on stacked encoders from a transformer. The new idea to transfer learning efficiently is relatively simple. First, a model is trained on unlabeled textual data which is in abundance. This results in a model trained to predict the following words in sentences and therefore understands the language quite well. BERT learns to understand the language by training on two unsupervised tasks: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). In short, to learn the language BERT breaks down the texts and tries to find patterns in the choice of words in sentences and the order of sentences. The second step is that this model can be fine-tuned for the task at hand by adding these last task-specific output layers by training on a supervised dataset. The significant advantage of this approach is that the initial model already has all the language understanding and you do not need a vast dataset tuned explicitly for your task. The heavy lifting is already done and with BERT this is done by Google. Finance is one of these niches that BERT is perfect for. The pre-trained BERT knew how to read but needed to be taught how to read like a financial analyst. To teach BERT this, a dataset consisting of financial texts was required and this dataset is Reuters TRC2 (Lewis et al. 2004). This dataset will train the model to understand the financial jargon found in financial documents. In addition to this domain training, a final dataset is needed to train the model for the sentiment analysis task. Luckily, this dataset is created and called the Financial Phrasebank (Malo et al. 2014). This small dataset consists of carefully labeled sentences extracted from various news articles, including financial statements. These sentences are labeled by experts in finance and master students multiple times. The final labels are presented with the inter-annotator agreement level for each sentence. A schematic overview of the different training steps is shown in Figure 3.

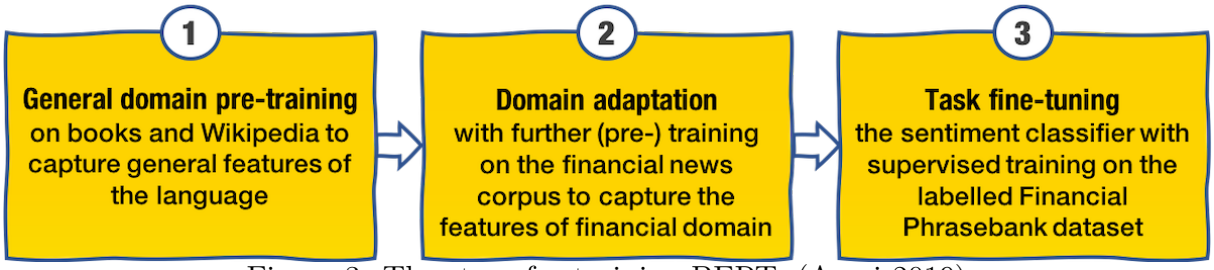

Figure 3: The steps for training BERT. (Araci 2019)

Because BERT is a transformer-based language model, making it a classification model is pretty straightforward. A classification output layer is added after BERT's initial layers which are used for sequential tasks like sentence classification or textual entailment. This final train step will result in FinBERT being fine-tuned to classify financial sentences based on sentiment in three categories positive, neutral or negative.

4.2 Textual analysis: Harvesting data

Now that the sentiment analysis model is trained the data needed for textual analysis can be harvested. The data used for the textual analysis consists of annual reports of companies related to the manufacturing and IT industry. Harvesting these yearly reports can be a tedious process and companies have started existing to make this process easier. In the United States (US), the SEC has created a platform called EDGAR which can be seen as a data warehouse for all related disclosures of companies in the Standard and Poor's 1000 index, better known as the S&P 1000. This thesis focuses on gaining insights into companies by investigating a firm's future state. The focus is on businesses operating in the US because these datasets are readily available and can be used to train and develop an algorithm that can later be used for annual reports found in different countries. The MD&A section consists of reports written by management that could contain valuable information about future possibilities or investments not known to the public. The EDGAR system includes these MD&A sections for the US market and these filings can be downloaded and filtered accordingly. The filtering is based on two criteria. The first criteria is that a focus will be made for this research only to include companies related to the three industries. The second filter is based on whether the companies have sufficient corresponding Compustat data needed for the model training. After filtering the harvested data will consist of MD&A filings dated between 1995 and 2022 of companies operating in the manufacturing and IT industry.

4.3 Textual analysis: Pre-processing

The pre-processing step is one of the most crucial steps in textual analysis. This step will make a computer be able to read humanized documents. Depending on the data type the preprocessing steps are very different. Luckily, lots of research is already done in this field and parsing pdf and XML files is implemented in multiple Python packages. Using a package like "Beautiful soup" and/or "XML parser" will result in clean text without tables, figures and styling. This result is now digestible by a computer and ready to be further preprocessed. For the sentiment analysis, the reports are needed with all the symbols, interpunction and capital letters as BERT is trained on these type of texts. However, for the choice of words model the texts need to be further processed and therefore the 10-K filings are parsed and cleaned into individual words where standard textual analysis conventions are followed. Common stop words such as *a, an, are, of, the* and *is* are excluded. In addition to this rule the following exclusions are also applied: single-character words, words not listed in the dictionary such as names firms and locations and all words that occur less than 5% in all documents. A name entity recognition (NER) algorithm is applied to check the list of words for any remaining words to be removed such as numbers written in words and wrongly classified names. For this the NER algorithm of Spacy is adjusted and implemented (Jiang, Banchs, and Li 2016). In Figure 4 the NER algorithm results are visualized. It can be seen that it can detect multiple types of words and in this approach only names, dates, organizations, cardinals and money are removed. Locations and Geopolitical Entity (GPE) are kept as it is found that these words contain helpful information for predicting a company's future state.

Figure 4: MD&A filing for Coca Cola Company (2011) NER visualized.

The last step of preprocessing is stemming the word list in such a way that the same variations of words are grouped. This is done to greatly reduce the number of words that need to be analyzed and therefore will decrease the number of variables. This will improve the results and significantly reduce the

computation time. Stemming is the process of reducing the word to its word stem that affixes to suffixes and prefixes or roots of words known as a lemma. In simple terms stemming is reducing a word to its base word or stem in such a way that the words of a similar kind lie under a common stem. For example, The terms care, cared and caring lie under the same stem 'care'. The words are stemmed using the snowball stemmer which is a stemming algorithm and it is also known as the Porter2 stemming algorithm as it is a better version of the Porter Stemmer (Porter 2001). The snowball stemmer is a more aggressive stemming algorithm and words are therefore more easily grouped. In addition to the stemming, in this model it is chosen to have the i's at the end of most stemmed words removed. This is done to ensure that words are correctly counted in the next step. For example, the word easily is stemmed to easili and in the following step the word would not be found in the text as it does not exist, whereas the word stem easil would.

4.4 Textual analysis: Analysis

The textual analysis will consist of two parts: Sentimental analysis and word-based analysis. Both will be used to try and predict critical metrics and/or indicators that can help predict the future state of a company.

4.4.1 Sentiment analysis

Sentimental analysis is something that has often been done in financial disclosures. Text sentiment is a concept taken from NLP literature. It can be defined as a measure to what extent the texts are positive or negative. It is used to classify texts and reflects the author's orientation concerning the content. Using the trained FinBERT model of Subsection 4.1 it is possible to classify sentences into three categories: Positive, Neutral and Negative. When a text contains more positive sentences than negative sentences it gets a more positive tone and vice versa. The following equation can capture the sentiment of text.

$$SI \; = \; S_{entiment-index} = \frac{S_{entences,Positive} - S_{entences,Negative}}{S_{entences,All}} \tag{1}$$

In this equation the neutral sentences will dilute the effect of the positive and negative sentences by increasing the denominator. The sentiment analysis will result in a number ranging between -1 and 1 which will indicate the tone of a text. This resulting number can be combined with the found factors to determine if the sentiment contains any additional predictive power. Because FinBERT understands language and not only words it can correctly extract the sentiment of sentences in texts. In contrary to a dictionary based approach, words are not stripped from their context. For example, the word growth is often classified as a positive word but growth in complaints is clearly not a positive development. The use of FinBERT does not rely on a dictionary which needs to be manually created. This creation is a time consuming process and one could argue that the word dictionary contains the subjectiveness of the creator.

4.4.2 Lasso regression

When text sentiment is correctly gauged the next part of the analysis can be conducted. In this part, the future state of a company will be predicted and therefore the total revenues and EBITDA are estimated. In textual analysis two approaches are often at the base of the study namely, relying on a word list or implementing a machine learning algorithm. The former has one significant disadvantage that it is very subjective because the researcher creates the dictionary and the latter is often criticized due to the "black-box" nature of most machine learning methods. In this research a new method is proposed which has high interpretability. A statistical machine learning approach that facilitates this interpretability is accomplished by combining a Lasso regression with a subsequent factor analysis (Tibshirani 1996). In this research, Lasso is chosen as a selection operator that can select words without any prior input and it can handle data that suffers from multicollinearity. It is expected that among the words this will occur. In addition, factor analysis is a dimension reduction technique and has good additional properties. In this situation it facilitates the increase of the interpretability by grouping words and creating interpretable factors. Therefore, combining both strengths, the selection power of Lasso with the interpretability of factor analysis, will result in a better understanding of which keywords may define the future state of a company. However, before lasso regression can be implemented the term frequency-inverse document frequency (TF-IDF) needs to be calculated for all the different annual reports. In short, TF-IDF is a measure, used in the fields of textual analysis, that can quantify the importance or relevance of string representations, in this case words, in a document amongst a collection of documents (also known as a corpus). An in depth explanation can be found in Section B of the appendix. An important thing to note is that to prevent data leakage between the training and test set it is crucial to keep the training and test set separate when calculating the TF-IDF. In past research this often goes wrong. This is because to calculate the TF-IDF the term frequency is scaled by the properties of the corpus but this may not contain information about the test set. Therefore when scaling you may only use the properties of the corpus consisting of the training set. The Lasso operator will create a unique subset of words whose frequencies best predict the firm's future state. The objective function for the one year ahead prediction is shown in equation 2 and in this also the penalty term can be seen. This penalty is what selects which words are considered insightful for predicting the future state. Due to the penalty no prior restriction is needed and therefore no subjectivity or prior beliefs are implemented.

$$(\hat{\alpha}, \hat{\beta}) = min \sum_{i,t} \left(revt_{i,t+1} - \alpha - \sum_{j} \beta_j \ freq_{i,t,j} \right)^2 \ s.t \ \sum_{j} |\beta_j| \leq C \tag{2}$$

In this equation, The predicted values are α, the constant coefficient, and β the freq coefficients. The firm's total revenues for the subsequent year are denoted by $revt_i, t+1$. For further predictions this value will be shifted even more. Researchers often scale the total revenues by the current end-of-year total assets to make the results more comparable among the observations. However, in this research it was found that this drastically increased the variance of the predictions and therefore was chosen not to implement this scaling (Eisfeldt and Papanikolaou 2014). The freq denotes the frequency of the j^{th}

word in the 10-K form of the company i in year t. Due to the penalty term, as mentioned earlier, many coefficients will be shrunken to zero. This variable selection still stabilizes the estimation (Tibshirani 1996). When implementing Lasso from packages constraint C is often chosen such that it minimizes the cross validation estimation errors. The same equation is used for the EBITDA estimation where $revt_i, t+1$ is substituted for $ebitda_i, t+1$. In addition, scaling is not needed as EBITDA is already scaled by a company's current assets. With this approach Lasso can find words that can help indicate the future state and help predict future company metrics. In other words, Lasso finds words that managers use, probably with no intended purpose, that help predict these variables.

4.4.3 Factor analysis

The above mentioned Lasso regression results in a dictionary of words with document frequencies that predict the future state. However, due to the high dimensionality the interpretability is low. By implementing a factor analysis the dimensions can be reduced and the results can be better interpreted. Two things need to be noted about the factor analysis. The first is that factor analysis is sensitive to outliers and the second is the assumption that assumes that each observed variable (word frequency) is a linear combination of some underlying latent factor and a normally distributed error term (Pett, Lackey, Sullivan, et al. 2003). The factors are estimated by exploiting the within correlations among these variables . Applying the famous Kaiser rule will only retain factors that contain more explanatory power than itself (Kaiser 1960). This combination of lasso and factor analysis has the following properties: ts training requires no additional datasets as annual reports and financial statements are widely available and the results are highly interpretable. This two-step procedure increases our understanding of the economic processes and their accompanying words which drive the future state of a company. Before applying factor analysis, highly correlated words are removed to ensure stability in the factor analysis. This is because factor analysis fails when highly correlated factors are included. In addition, Bartlett's test of Sphericity and a Kaiser-Meyer-Olkin test will be applied (Tobias and Carlson 1969)(Dziuban and Shirkey 1974). The former is to ensure that the correlation matrix is not an identity matrix, meaning that the variables are unrelated and not ideal for factor analysis. The latter is a test conducted to examine the strength of the partial correlation (how the factors explain each other) between the variables. These tests are further discussed in Section C in the appendix. Furthermore, the principal component method is used to identify the factors (Passamani, Tamborini, and Tomaselli 2015). As commonly done, the priors required by the factor analysis procedure are set to the squared multiple correlations (SMC) which can be interpreted as how much the other observed variables explain each observed variable. After the factor analysis a promax rotation is applied to the standardized, un-rotated factors to allow for correlation between them. This increases interpretability and performance because it reduces the number of variables (Pett, Lackey, Sullivan, et al. 2003). In addition to this reduction, the remaining factors are interpretable because they can be labeled based on words with high loadings and high scoring observations.

4.5 Prediction

After interpreting the factors found in the previous step it is time to evaluate the power of the found factors in predicting the future state of a company by predicting the total revenues and EBITDA for the three different timeframes. A regression will be done on the company metrics, regressing upon the found factors and the sentiment index (SI). One could stand to reason that extra steps should be taken to account for the COVID-19 pandemic. However, after running multiple tests it was found that accounting for this event did not improve the results and often resulted in worse predictions. Two reasons for this are because the company metrics of the companies in the S&P 1000 remained quite constant and for the used time frame the covid crisis had not peaked yet in terms of revenue decline. In-sample regressions will be made on the training set to evaluate the informativeness of the variables. This is done by calculating the incremental R^2 of the variables. The found in-sample coefficients help interpret to what extent they account for information that can indicate the future state of a company. The out-of-sample predictive ability of the factors is assessed using the Pseudo R^2. The in-sample regression coefficients are used to predict the future state of a company for the test set to estimate the mean absolute error (MAE) and mean squared error (MSE). In this paper the root of the MSE will often be presented (RMSE). This is done because the term seems to explode due to the high variation of different companies in the dataset. For more clarification on the definitions see Section D in the appendix. The out-of-sample pseudo R^2 is calculated as one minus the ratio of MSE from a forecasting model to that of a model with only the intercept included (Campbell and Thompson 2008). This benchmark is valid because it captures the predictive power of the historical means of the dependent variables. Since revenues are sticky and often not volatile, the historical means typically predict these future realizations well. The equations and more information about this benchmark can be found in Section D of the Appendix.

5 Data

For both the sentiment analysis and the lasso-factor analysis the annual reports of US-listed companies (10-K) retrieved from the EDGAR database operated by the SEC will be used. By filtering the Standard Industry Classification (SIC) it is possible to select only companies related to specific industries. The SIC codes used can be found in the appendix in section A. This selection will provide a list of all companies that have published annual reports, up until 1995, in these categories. Specifically, the data set will contain: company names, state/country identification codes, company identification numbers and central index keys (CIK) which are needed to link them to the corresponding Compustat data. All companies need to be checked for sufficient data as insufficient data will result in the corresponding observation removed. After thorough literature research it is chosen that the focus of this research will be on the MD&A section of the 10-K's. These items 7 & 7A of a 10-K form, are shown in the schematic breakdown of a 10-K form shown in Table 1.

Table 1: 10-K table of contents.

section		**Description**
PART 1	item 1.	Business
	item 1A.	Risk Factors
	item 1B.	Unreserved 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 Issue 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 and Financial Statements Schedules

Source: SEC EDGAR website: https/www.sec.gov/edgar/searchedgar/companysearch.html

The sentiment analysis is conducted for all observations. In contrary to the lasso model which requires a training sample of 1995 - 2017. It is chosen that 2018 - 2020 is left out as a test set. For the lasso model further preprocessing was needed as mentioned in section 4. These conventions significantly reduce the number of words in each text and make this analysis much more feasible. These words are reduced to their initial roots called word stems. The TF-IDF of every unique word stem is calculated and lasso selects the stems whose frequencies predict the future state of a company for multiple time frames. After factor analysis, the various dependant variables are regressed on the found factors with and without SI. The found coefficients are used to calculate and predict the values of the dependent variables in our test sample. The baseline regressions are compared with prediction models already used in finance to test the incremental explanatory power. To indicate the out-of-sample predictive power the pseudo R^2 is calculated using the actual and predicted values of the dependant variables in the test sample.

6 Results

In this section, the results will be presented and discussed. The two different company metrics, total revenues and EBITDA, will separately be discussed. Both metrics are predicted in similar ways and therefore only the total revenues will be thoroughly discussed. The EBITDA analysis will be briefly presented for the manufacturing industry in the last part of this section.

The total revenue prediction results are discussed for three different time frames namely, one, three and five years ahead for several different models. The results for the manufacturing industry will be extensively discussed whereas the results of the IT industry will briefly be presented as the analysis is very similar and these results are merely added for validation and comparison. This sections consist of four parts where every part focuses on a different aspect of the model.

The first part consists of presenting and discussing the factor analysis results. These factors can be used for predicting the total revenues but more information and insights can be extracted. The second part will focus on the results of the sentiment analysis and the question if adding the sentiment index would improve the model's performance. A comparison will be made between two different models to see if the sentiment analysis extracts additional useful information to predict a company's future revenue. The third part will compare three different models to validate their predictive powers. The three models will predict the total revenues of the test set and will be compared afterwards. The first model consists of only the found factors whereas the second model will have the SI added. The third model will predict total revenues based on a prediction method that is often used in finance. A model consisting of only the SI will not be discussed as it was found in untabulated tests that it did not perform well. The last part will be focused on outliers in the dataset. Extensive analysis will be performed on the one year ahead manufacturing model to investigate the robustness properties of the proposed approach. It was found in untabulated tests that the outlier analysis gave similar results for the other time frames and the IT industry.

6.1 Factor analysis

Table 2: Overview of the results for predicting future total revenues.

Forecast:	**# Observations:**		**Lasso:**	**Factor Analysis:**		**R^2:**	
Manufacturing	**Training**	**Test**	**Selected Words**	**# Factors**	**$Var_{exp-test}$**	**Training**	**Test**
one year	2175	292	875	184	94%	0.94	0.93
three years	1778	281	930	186	92%	0.92	0.82
five years	1438	261	28	10	80%	0.80	0.81
IT	**Training**	**Test**	**Selected Words**	**# Factors**	**$Var_{exp-test}$**	**Training**	**Test**
one year	5486	1033	198	62	76%	0.75	0.60
three years	4019	807	157	50	71%	0.70	0.61
five years	3017	689	151	10	65%	0.64	0.56

Notes: For the training results only the factors are used whereas for all test set results the complete model (factors + SI) is implemented. $Var_{exp-test}$ is the explained test set variance of the model. The R^2_{adj} is not displayed as it was pretty similar to the R^2 which indicates that the factor analysis has done an excellent job in reducing the number of variables.

Three forecasts for the manufacturing and IT industry are made for three different time frames. Each forecast results in a different set of factors. The found factors are ranked based on their incremental R^2. The incremental R^2 is the resulting decrease of the R^2 when only that factor is removed from the regression of the training set. For a more extensive explanation see Section D in the Appendix. For the manufacturing industry only the first five factors are thoroughly discussed to conserve space and the labels are given for the top 15 factors. For the IT industry only the labels are presented. It was found in all regressions that the top 5 factors are positively correlated with future total revenues. Although not all factors are discussed in this section they are all included in the presented regressions of the Results section.

6.1.1 Manufacturing - Factors

The Manufacturing dataset contains 2467 annual reports dated between 1995 and 2021 consisting of 300 different companies. The training/test split is at the beginning of the fiscal year 2018. Due to sparse data, increasing the forecasted time frame results in a decrease in the size of the training and test set. After preprocessing the data 1917 unique word stems are found. A summary of the prediction results is shown in Table 2. The found factors are schematically shown in Table 3. In Tables 12, 13 and 14 in Section E of the Appendix are the top 5 factors displayed along with the top ten stemmed words with the highest factor loadings and the 20 highest scoring observations (Firm - year).

Table 3: Factor summary and labels of factors 1 till 15.

	Lasso	**Factor Analysis**	**Rank Factor**					
Forecast	**# words**	**# factors**	**1**	**2**	**3**	**4**	**5**	**6 till 15**
one year	875	184	corporate structure	operations	company governance	finance	beverages	health, product portfolio, seasonal, dairy, brand, market, legal, coffee, sourcing, consumer and markets
three years	930	186	beverages	fair trade	supply chain	finance	retail	workplace environment, well being of employees, seasonal products, taxes, Mexico, delivery channels, penetration, real estate, customer, retention, organic and farming
five years	28	10	products	commodities	dairy	beverages	expansion	company health, sourcing, transport, expansion, organic

For the one year ahead forecast, lasso is applied on the found unique word stems. It selects 875 word stems whose frequencies best predict the total revenue. When the adequacy of factor analysis is ensured

it can be applied to the remaining frequencies which result in 184 factors whose eigenvalues exceed one and account for 94% of the variance in the training set. The top five factors will be discussed below.
Factor 1, corporate structure is associated with the structure of a company and its leadership, how orders are delegated throughout the company. The 10-K's are often about how a company is to be restructured or the effects of a restructuring in the past. Words with high loadings which frequently occur in these 10-K's are staff, synergy, input, cyclic and split. Manufacturing firms from the food & beverages industry are associated with this factor above average.
Factor 173, operations is associated with how a company, specifically a factory, is run. When this is openly discussed in an annual report it often means a change has been made in the past or is scheduled in the future. Words with high loading for this factor are spoilage, workers, error, meal and lag. The observations that score high on this factor are often companies in the food manufacturing business.
Factor 3, company governance is associated with how a company is governed and what policies and benefits are in place or will be implemented in the foreseeable future. The words with high loadings related to these topics are allocations, medical, divisions and digitisation. High scoring firms are often older and more stable companies such as Pepsico Inc can be seen in the top firm-year observations.
Factor 0, finance is associated with the firm's financial state. The top words in these factors are impairment, tangible, assets, equity and analysis. The financial situation is an essential topic in a company's annual report. This results in that not one type of manufacturing firm stands out for this factor and all observations score adequately.
Factor 4, beverages is related to the processes of the beverage industry. All top scoring observations are of the big beverage firms. Words with high loading are bottles, case, syrup, concentrates, measured and competition. Competition is one of these high loading words because consumers are often not picky in the beverage industry. Therefore, competition in this industry is high and comparing with competing companies is the only suitable measure to see how much real growth is realised.

For the three year ahead forecast lasso selects 930 word stems whose frequencies best predict the total revenue. The subsequent factor analysis results in 186 factors whose eigenvalues exceed one and account for 92% of the variance in the test set. The first five factors will be explained in depth below.
Factor 1, beverages are once more related to the processes of the beverages industry. All top scoring observations are of the big beverage firms. Words with a high loading are bottles, case, leadership, concentrates, measured and deconsolidated. The factor seems to be the same as for the one year ahead forecast only that the words and corresponding observations have changed. The same companies appear, but in different years. Furthermore, the word leadership has become significant.
Factor 18, fair trade is about the product's origin and how the materials are harvested. In addition, fair trade is also about the trust a consumer can have in a company. Do the consumers believe the story that the company tells is the truth. Words with high loadings are cocoa, obligations, trust, supply chain and vote. Very different companies score high on this factor ranging from alcohol manufacturers to clothing brands.

Factor 3, supply chain is associated with manufacturing company's sourcing and supply chain. This is a vital internal part of a manufacturing company because they often cease to exist without a solid supply chain. Words that are affiliated with this factor are commodities, oilseed, counterparties, wheat and crop. The factor seems to be dominated by the company Bunge LTD but looking at more observations it can be seen that also other companies score high on this factor. It can be stated that the companies often do reside in the agricultural manufacturing business.
Factor 0, finance is once again associated with the firm's financial state. The top words in these factors are pension, foreign, equity and goodwill. When comparing the words of this factor with the finance from the one year ahead prediction, many similarities can be found like equity, methods and derivatives. However, it seems as if employee welfare has become more critical because words like pension and goodwill have increased their loadings for this factor. Observations with a high score for this finance factor are often more mature companies.
Factor 4, retail is about where and how a consumer buys the manufactured products. This factor includes online retail as well and words affiliated with this factor are openings, stores, websites, merchandising and wholesale. Companies that score high on this factor are clothing brands but average scoring companies can also be found in the consumer products industry.

For the five year ahead forecast lasso is applied on 1873 unique word stems. It selects 28 words whose frequencies best predict the total revenue. Factor analysis on the remaining frequencies results in 10 factors whose eigenvalues exceed one and account for 80% of the variance in the training set. The first five factors will be thoroughly explained below.
Factor 2, products is about the portfolio of products a company sells or manufactures. High scoring observations on this factor are companies discussing a change in their product portfolio. Words with high loading are incentive, products, cheese, expenses, increases and increments. Companies that score high in this industry are often manufacturers for the fast moving consumer goods industry.
Factor 1, commodities is about the sourcing and materials needed to manufacture products. These materials often come from less developed places in the world. Words with high loadings are commodities, agricultural, grains, livestock and land. Large food processing companies are high scoring observations.
Factor 5, dairy is about products made from farm animals, and high scoring observations are all in the food production industry. Words like cheese, milk, grain, and Asia have high factor loadings.
Factor 0, beverages are still the factor affiliated with the beverages industry. However, for the five year ahead forecast it can be noted that different words have higher loadings. Words with higher loadings are more supply chain related like Africa, middleman and digitise.
Factor 4, expansion is associated with the growth of companies and restructuring their internationally oriented supply chain or consumer focus. Words with high factor loadings are metrics/KPIs, translation, stronger, global and Asia. This factor is dominated by clothing brands and textile producers for the manufacturing industry. A word affiliated with this industry is selling, general, and administrative expenses (SG&A) which is often discussed in their MD&A section.

6.1.2 Manufacturing - Factor comparison

A few things can be noted when comparing the factor analysis results for the different time frames. Firstly, it can be seen that some factors seem to be stable for the different time frames like finance and supply chain. These factors remain in the top 15 and this stands to reason that these factors are essential in smaller and larger time frames. These are often vital components of a company and one can understand why these factors remain important for the different time frames. In contrary, some factors like commodities and products become only significant when looking at a larger time frame. Their incremental R^2 increases when predicting total revenues further in the future. It stands to reason that predicting a company's revenue further in the future will result in word frequencies that reflect on the future becoming more important.

Focusing on the observations associated with different factors new insights can be gained. It can be seen that observations with a high score associated with various factors give insights into their primary focus at that particular moment in time. A high scoring observation means the annual report contains those words with high factor loadings. This is valuable information because it can give an indication to what is the content of a yearly report and the focus of the company for that fiscal year.

Looking at the years that different factors have become important for different observations one can extract information about the state of the market for that industry. For the manufacturing industry one of these notable events is the financial crisis of 2009. Factors reflecting on the future and restructuring become increasingly crucial around that period.

Lastly, exploring all the chosen factors one can see that two different types of factors are found. Company specific factors and industry specific factors. industry specific factors can be used to categorise these companies in more specific segments in their corresponding industry.

6.1.3 IT - Factors

For the IT industry three forecasts are made for the different time frames. The IT dataset contains 6519 annual reports dated between 1995 and 2021 consisting of 1030 different companies. After preprocessing the data 1909 unique word stems are found. A summary of the results is shown in Table 2. The first ten factors for the three different time frames are presented in Table 4. As was done for the manufacturing industry, the top 5 factors are presented along with the top ten stemmed words with the highest factor loadings and the 20 highest scoring observations (Firm - year) in Table 15, 16 and 17 in Section F of the Appendix.

Table 4: Factor labels of factors 1 till 10.

Forecast	Factor labels
one year	interactive, corporate structure, finance, cloud, service, production, sourcing, taxes, car manufacturing and medical
three years	cost and expenses, marketing, restructuring, electronic arts, cloud, japan, advertisement, outsourcing, merger and subscription model
five years	strategy, expenses, business model, revenue, operations, communication, software, distribution, outsourcing, franchise and title oriented

The labels of the top ten factors are schematically shown and it can be seen that the results are very similar to the manufacturing results. Altering the forecasting length changes the importance of different factors and increasing the forecasting length results in future oriented factors becoming more important. The model selects other words which often have different future orientations. Comparing the found factors between the industries it can be seen that some factors remain in the top 5 for both industries. These factors are often more general and critical to a company's revenue. These factors are corporate structure, finance, marketing and strategy. In contradiction to the manufacturing results one can see that industry specific factors, like car manufacturing and dairy, behave differently. For the IT industry these factors become less important in larger time frames whereas for the manufacturing industry they remain important. It stands to reason that this is because the manufacturing industry has well defined segments which have not drastically changed these last couple of decades. At the same time, the IT industry has changed a lot and is still evolving. The IT industry is growing at a rapid pace and therefore it seems that industry specific segments do not behave in the same way as the well defined industry specific segments. One thing to add is that companies in the IT industry often flow with the market's direction and change their operations accordingly which could further deviate the behaviour of these industry specific factors.

6.2 Sentiment analysis

The sentiment analysis is by far the most time consuming process of the model. It increases the computation time by a factor of 100 compared to the choice of words model. This is because running annual reports through the sentiment model requires much computation time. This is due to the nature of BERT and its methods for understanding text. The sentence focus of BERT has its advantages but is also very computationally intensive. Because the sentiment analysis is trained on sentences this leads to the fact that for every text it analyses, it first breaks the text down into sentences and then applies sentiment analysis on all the different sentences, one by one. In this section, the sentiment analysis results will be presented for the three different time frames and two industries. The analysis will be done for two models. The first model will consist only of the found factors whereas the second model will include the SI and the factors. The results for the training and test set will be presented using the R^2, Pseudo R^2 and incremental R^2. In addition, the significance (5%) of the sentiment coefficients , MAE and RMSE are shown. With all these results it is possible to conclude if adding the SI to the choice of words model will indeed improve predictions.

The results of the manufacturing industry analysis will be presented first and thoroughly discussed, whereas the IT industry analysis results will be stated and briefly discussed. Afterwards, these results will be compared.

6.2.1 Manufacturing - Sentiment

Table 5: Results of factor only model and factor + SI model.

	Training		Test		
Model (one year)	**R^2**	**SI Significant**	**R^2**	**MAE**	**RMSE**
Factors	0.93	-	0.93	2159	3173
Factors + SI	0.93	Yes	0.93	2114	3114
Model (three year)					
Factors	0.94	-	0.87	1519	2298
Factors + SI	0.92	No	0.82	2332	3493
Model (five year)					
Factors	0.945	-	0.87	1488	2175
Factors + SI	0.80	Yes	0.81	3153	5398

In Table 5, the results are shown for the two models on the three different forecasting lengths for the manufacturing industry. The model performs well on the training set and this was expected because the factors are derived from the words that best predict the total revenues in the training set. The in-sample regression coefficients are used to help interpret how the factors predict total revenues. The model is built so that no data leakage could have happened between the training and test set. The Mean squared error (MSE) is calculated and from this the out-of-sample Pseudo R^2 is calculated to asses the found factors (Campbell and Thompson 2008).

The incremental R^2 for the SI index for the forecast one year, three years and five years ahead are 0.0013, 0.00015 and 0.0037 respectively. From the results of the regression it can be seen that the SI is positively

correlated with future total revenues. When looking at the table it can be seen that predictions in the foreseeable future do indeed benefit from adding the sentiment. However, A slight decrease in errors can be noted. The model's performance with the SI included starts to decrease when the prediction length starts to increase. This leads to worse predictions as a lot of variation is added into the model. With backwards reasoning this feels intuitive. A company's far future state should not depend on the sentiment at this moment. Periods where sentiment is low would then create future periods where the total revenues would decrease. However, looking at a shorter time frame it stands to reason that positive sentiment in an annual report could mean that the company is financially stable and perhaps improving.

6.2.2 IT - Sentiment

Table 6: Results of factor only model and factor + SI model.

	Training		Test		
Model (one year)	**R^2**	**SI Significant**	**R^2**	**MAE**	**RMSE**
Factors	0.73	-	0.59	876	2047
Factors + SI	0.75	Yes	0.60	862	2027
Model (three year)					
Factors	0.69	-	0.65	634	1379
Factors + SI	0.70	No	0.61	680	1541
Model (five year)					
Factors	0.60	-	0.63	717	1573
Factors + SI	0.64	No	0.56	704	1773

In the table above the IT industry results are shown and similarities can be seen when compared with the manufacturing results. The incremental R^2 is relatively low for the SI index in all time frames and positively correlated with future total revenues. The addition of the SI to the model has a positive effect on the prediction accuracy for the small time frame predictions but a more significant negative effect on the larger time frame predictions. A slight difference is that for the five year ahead prediction the SI was insignificant whereas for the manufacturing industry it was significant. The results of the analysis seem to behave similarly for both industries. The change in performance for the different models and prediction time frames is similar. In other words, the models change analogous when altering the forecasting lengths.

It can be noted that the IT industry's overall results are significantly lower than the manufacturing industry for both models. The R^2 of the models are lower for the IT industry than for the manufacturing industry. This means that the model does a bad job of replicating the observations. This can have numerous reasons but in comparison with the manufacturing industry one difference stands out. The IT industry has changed a lot these past decades whereas the manufacturing industry has not. Different types of IT companies are created and the industry becomes broader and inherently more different. These companies often have other goals, a different focus and a different view of their market position. Combining all these different companies and their annual reports perhaps generalises these companies too much. To give an illustration of the change in the two industries. The S&P 1000 contained 79 IT related companies and 234 manufacturing related companies in 1995. Whereas in 2020 it held 572 IT

related companies and 283 manufacturing related companies. In addition, during these times lots of IT companies have entered and exited the S&P 1000. It stands to reason that similar companies have similar goals and that these companies report similar matters. The model benefits from these similarities and uses this to make better predictions. If the companies in a dataset become inherently more different then the model starts performing worse. In Section 7, a different approach is proposed which could improve the model for industries with varying types of companies.

6.3 Validating the model

In this section, three different models will be compared with each other. The first model contains only the found factors. The second model includes the found factors and the SI and the third model is a prediction based on fitting a polynomial through past total revenues which is often done in finance (Grizzle and Klay 1994). One thing to note with this method is that company data is sparse and a limitation for a company to be included is necessary. Only companies with data points in the test period and more than three training data points are included. This is to ensure that the third model does not extrapolate from two points. Therefore, the test set size will decrease and the first two models need to be adjusted based on this new smaller test set to still be correctly compared. This is done for all three different forecasting time frames. The manufacturing industry will extensively be discussed and afterwards, the results of the IT industry will be presented and briefly elaborated as is done in previous sections.

6.3.1 Manufacturing - Validation

Table 7: Comparison of the three models.

	Test			
Model (one year)	**Size**	**R^2**	**MAE**	**RMSE**
Factors	228	0.94	2215	3262
Factors + SI	228	0.93	2198	3201
Polynomial Fit	228	0.94	1147	3201
Model (three year)				
Factors	206	0.87	1467	2028
Factors + SI	206	0.73	2052	2998
Polynomial Fit	206	0.85	804	2210
Model (five year)				
Factors	197	0.88	1374	1962
Factors + SI	197	0.40	2859	4410
Polynomial Fit	197	0.84	771	2092

In this table the comparison of the three different models can be seen. It can be seen that for the one year ahead predictions all models score quite similar. The model with the SI included performs slightly better than the factor only model. When increasing the prediction time frame to three years ahead the test set decreases in size as was expected. It can also be seen that the model with the SI included has a greater decrease in performance in comparison with the other models. In the largest time frame it can

be seen that the SI model does not perform adequate anymore. The other models still perform quite well and it can be seen that the factor only model starts performing better than the polynomial fit.

6.3.2 IT - Validation

Table 8: Comparison of the three models.

		Test		
Model (one year)	**Size**	**R^2**	**MAE**	**RMSE**
Factors	553	0.58	2739	5923
Factors + SI	553	0.57	2729	6148
Polynomial Fit	553	0.81	924	4189
Model (three year)				
Factors	476	0.54	1172	3225
Factors + SI	476	0.46	1175	4528
Polynomial Fit	476	0.70	804	2477
Model (five year)				
Factors	447	0.51	1224	2120
Factors + SI	447	0.32	1224	7180
Polynomial Fit	447	0.62	1100	1800

The results of the IT industry show similar patterns with the manufacturing industry. One can see that the addition of the SI drastically decreases the performance of the predictions on longer time frames as was found for the manufacturing industry. In addition, the decrease of the prediction accuracy is more rapid for the financial predictions as for the factor model. However, for the IT industry the model does not outperform the extrapolation model. This could be a consequence of what can be seen in Table 6. The initial training R^2 is quite low which means that the model does not capture enough information to make adequate predictions for the training set which often results in lower R^2 for the test set. One big difference with the manufacturing industry is the number of different companies the dataset consist of as mentioned in Subsection 6.2.2. These different companies can eventually result in the model having trouble finding sets of words which can predict future revenues for these different companies. One thing to note is that this approach does not utilize the information that distinguishes these different companies. It sees every observation as a unique observation. This approach does work for the manufacturing industry because companies are more similar and more stable over time. However, the performance worsens when different types of companies are combined. A solution for this problem is proposed in Section 7.

6.4 Outlier Analysis

In this section the data will be analysed to see if outliers are present in the data and if so what type of outliers. This will be done for the one year ahead predictions of the manufacturing industry as similar results were found for the IT industry and the other forecasts. As mentioned in Subsection 4.4.3, factor analysis is found to be not robust. This is because in factor analysis, one needs to estimate the matrix of factor loadings (Λ) (which is only specified up to an orthogonal transformation) and a diagonal matrix containing on its diagonal the specific variances Ψ. Classical factor analysis methods are very vulnerable to the presence of outliers. Luckily, methods are constructed which can resist the effect of the different types of outliers.

In classical factor analysis, the matrix $\mathbf{\Sigma}$ is estimated by the sample covariance matrix estimator. Afterwards this is decomposed to obtain the estimators for Λ and Ψ. Many methods have been proposed to improve this decomposition, of which maximum likelihood (ML) and the principal factor analysis (PFA) method are most frequently used.

In this research the minres implementation is used which is a combination of the ML and PFA approach. However, outliers can heavily influence the estimation of the variance matrix $\mathbf{\Sigma}$ and hence also the estimation of the parameters. Therefore it is better to implement a different type of estimator, one that is robust. One of these estimators is the Minimum Covariance Determinant (MCD) (Rousseeuw 1985). The MCD looks for the subset of h out of all n observations having the smallest determinant of its covariance matrix (typically, $h \approx 3n/4$). The MCD estimator is highly robust, has good efficiency properties and is available in several software packages. One limitation of the MCD-based approach is that the sample size n needs to be bigger than the number of variables p. This criteria is often met in practice. Recently, a fast MCD algorithm has been developed which in practice is often used (Rousseeuw and Driessen 1999). However, This is not implemented in this research due to time constraints. In exchange, a thorough outlier analysis will be conducted and it will be shown that for these datasets the outliers do not heavily effect the results.

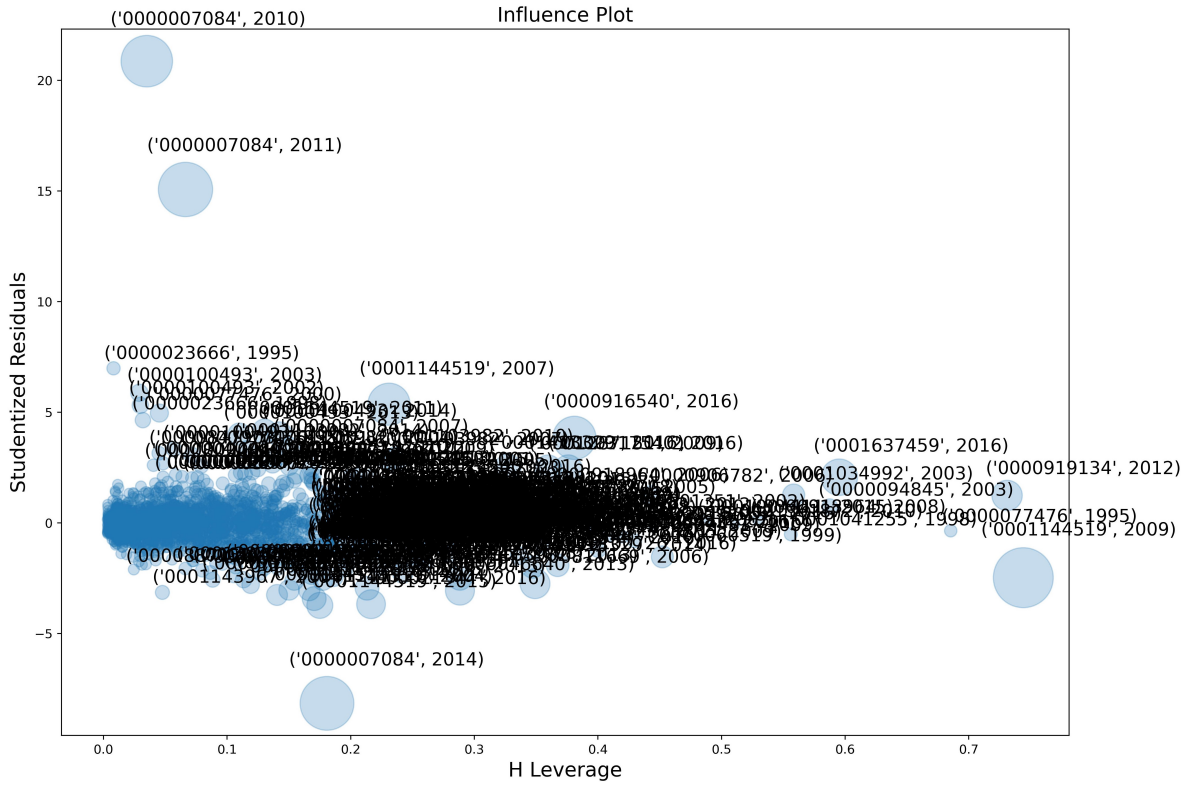

Figure 5: Influence plot of observations in the training set (1-year ahead forecast).

In Figure 5, all observations are plotted with leverage on the horizontal axis and influence on the vertical axis. The size of the points are calculated using cooks distance. In statistics, Cook's distance is a commonly used estimate of the influence of a data point when performing a least-squares regression analysis (McDonald 2002). As can be seen in the figure our model detects that there are outliers present in the data. Outliers with large studentized residuals are bad leverage points and they become worse when their leverage increases whereas outliers with high leverage but small studentized residuals are good leverage points which can even improve predictions. In Figure 6, the observations are plotted per year to see if outliers coincide with a specific time.

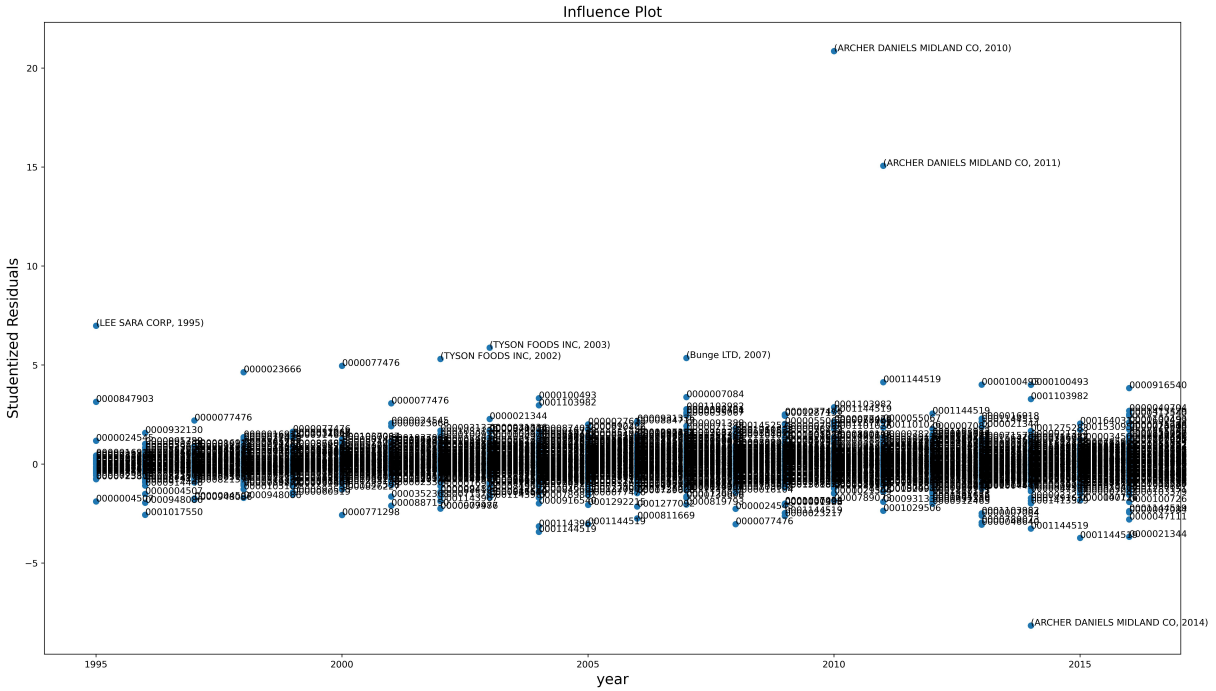

Figure 6: plot of observations in the training set (1-year ahead forecast).

In this figure it is not clear if a specific time or a certain company results in outliers. One thing that can be noted is that observations with large studentized residuals are often found in the last five years. To detect outliers, one of the things that is often done is performing a Bonferroni Outlier Test. The Bonferroni Outlier Test uses a t-distribution to test whether the models largest studentized residual values is statistically different from the other observations in the model. A significant p-value indicates an extreme outlier that warrants further examination(Kaalund et al. 2014). This test is implemented on the observations and on a 5% level, 9 observations are detected as outliers. To show that in this research the outliers do not have any drastic effects on the results all observations of companies with an outlier observation are removed. It was found that for the one year ahead forecast the 9 outliers resulted in 87 observations to be removed. The results of the analysis on the cleaned data can be found in Table 9.

Table 9: Comparison of the models with and without outliers.

Model (one year)	Training		Test		
With Outliers	**R^2**	**SI Significant**	**R^2**	**MAE**	**RMSE**
Factors	0.93	-	0.93	2159	3173
Factors + SI	0.94	Yes	0.93	2114	3144
Without Outliers					
Factors	0.94	-	0.87	1488	2175
Factors + SI	0.95	No	0.88	1393	2185

As can be seen in the table, removal of the outliers has minor effect on the the prediction power. The MSE has decreased because observations with high residuals are removed. A thing to note is that good leverage points are also removed which result in the R^2 decreasing as well. It is a shame that this is not the correct way to handle outliers as these are real verified observations and therefore should still be included in the data. A better solution would be to implement a robust version of the factor analysis by implementing the MCD estimator as mentioned earlier. However, Table 9 shows that not accounting for the outliers in this research does not have a large effect and the results remain valid.

6.5 EBITDA

In this section the EBITDA metric will be predicted for three time frames for the manufacturing industry. This analysis is added to show that the model is not limited to predicting total revenues and can be implemented to predict many different metrics, in this case EBITDA. This subsection is structured very similar to the total revenues result section. First, the added value of the SI will be investigated by comparing two models, a factor only model and a model with the SI and factors included. In addition, these results will be compared with the results found for the sentiment analysis of the total revenue predictions. Secondly, the two models will be compared with a model where EBITDA is extrapolated from historical data. In Tables 18, 19 and 20 of Section G of the Appendix, are the top 5 factors displayed along with the top ten stemmed words with the highest factor loadings and the 20 highest scoring observations (Firm - year). These factors will not be discussed as this is very similar to Section 6.1.

Table 10: Results of factor only model and factor + SI model.

	Training		Test		
Model (one year)	**R^2**	**SI Significant**	**R^2**	**MAE**	**RMSE**
Factors	0.95	-	0.91	446	770
Factors + SI	0.95	Yes	0.91	414	743
Model (three year)					
Factors	0.93	-	0.85	277	526
Factors + SI	0.91	No	0.83	311	552
Model (five year)					
Factors	0.87	-	0.78	331	646
Factors + SI	0.84	No	0.75	343	702

The comparison between the two models can be seen in Table 10. These results for predicting the EBITDA are quite similar to the results seen in Section 6.2. Short predictions benefit from adding the SI with a slight improvement in the prediction accuracy. However, further predictions perform worse when the SI is included. This is very similar to the results of the total revenues predictions. One thing to note is that in the predictions of the EBITDA, it does seem as the SI influence on the results is smaller. The two models do not deviate as much as was seen in the total revenue comparison. Another thing to note is that the RMSE is significantly lower than seen in Table 5. This is expected because the EBITDA score spread is much lower as it is scaled by a company's total assets.

Table 11: Comparison of the three models.

		Test		
Model (one year)	**Size**	**R^2**	**MAE**	**RMSE**
Factors	228	0.92	429	674
Factors + SI	228	0.92	426	672
Polynomial Fit	228	0.93	306	590
Model (three year)				
Factors	204	0.84	271	460
Factors + SI	204	0.74	370	510
Polynomial Fit	204	0.89	201	340
Model (five year)				
Factors	447	0.82	246	454
Factors + SI	447	0.59	420	640
Polynomial Fit	447	0.81	290	512

In this table the results of the three different models can be found. Comparing these results with the results of the prediction of the total revenues for the manufacturing industry, Table 7, not a significant difference can be seen but a lot of similarities can be found. Similarities that can be seen are that the sentiment model performance worsens when further predictions are made. In addition, the two other models score equivalent on the one year ahead predictions and both performances decrease when increasing the prediction time frame. The factor model decrease is slightly greater. The scores are equivalent for the five year ahead predictions with the factor only model performing slightly better.
What can be concluded from these results is that the model performs quite similar for different company metrics. In addition, the results of the EBITDA predictions and the total revenues predictions show analogous behaviour. These metrics are important company metrics and their fundamentals are discussed in the MD&A filings. It is for this reason that the model can predict these different metrics as it can be trained to find those sets of word frequencies which best predict the chosen metric. This is very noticeable when comparing the factor tables in the Appendix. The factors often show similar labels but the underlying words that define these factors are very different.

7 Conclusion

In this research, textual analysis was applied to annual reports of companies operating in the US manufacturing and IT industry. Annual reports dated between 1995 and 2020 were extracted from the EDGAR database. This was done to answer the question if the texts found in yearly reports contained useful and quantifiable information which could be used to improve or perhaps substitute financial prediction methods. After thorough literature research it was found that the focus would be on MD&A filings as they contained most information useful for predictions. In addition, the scope of the type of information to extract was limited to sentiment and a choice of words model. The sentiment analysis was done by implementing FinBERT and the choice of words model was realised by a lasso regression with subsequent factor analysis.

It was found that the choice of words model can extract quantifiable information which resulted to be useful for predicting company metrics. Multiple financial predictions can improve by implementing such a model as will be elaborated below. First, the results of the choice of words approach and sentiment analysis will be discussed. Secondly, the comparison between the three models will be presented and reflected upon. Thirdly, the improvements will be stated and future recommendations will be proposed.

The choice of words model was chosen because leadership writes the MD&A filings of a company. There is reason to believe that these filings contain useful information about the future state of a company. Managers choose their words wisely and perhaps unknowingly change their words when company prospects differ. The results of the choice of words model are interpretable factors. Besides their predictive power, these factors give new insights into what important topics are discussed in the MD&A section of a company. These factors can indicate a companies focus at a particular time, the important topics in a industry during a specific period and more.

The sentiment analysis was trained to understand financial documents and extract the sentiment of the writer, in this case the manager of a company. One could argue that a positive sentiment could perhaps indicate a brighter and better future. The sentiment analysis result was a negative, neutral or positive index for the MD&A filing. It was found for all short predictions, one year ahead, adding the SI in the model resulted in a slight improvement of the R^2 ($\approx 2\%$). Therefore, it can be said that the sentiment captured additional information in the text, making the predictions more accurate for short term predictions. However, further predictions resulted in far worse results and drastically increased the variation in the model. Combining these facts, it is perhaps better not to include the SI as this increases the amount of variance in the model only to gain a slight increase in accuracy in short term prediction and a much greater decrease in further predictions.

To answer the question of whether the used approach could improve or substitute financial predictions, the comparison of the three models predicting total revenues will be discussed. As mentioned above, adding the SI to the choice of words model often resulted in worse predictions. Therefore, only the comparison between the polynomial fit and the choice of words model will be discussed.

For the one year ahead predictions, both industries gave different results. For manufacturing industry all models gave comparable results where the polynomial fit seemed to perform slightly better. However

for IT, the polynomial fit performed adequate whereas the other models performed quite lousy. When increasing the prediction time frames all models for both industries behaved similar. The largest decrease in performance was found for the SI included model. A slightly less decrease was seen in the polynomial fit and even less was the decrease of the factor only model. For the manufacturing industry, the factor model started performing significantly better than the polynomial fit for the larger time frame predictions. For the three years ahead forecast the R^2 improved by 2% and the RMSE decreased by 8% and for the five year ahead predictions the R^2 improved by 4% and the RMSE decreased by 6%. This result was not seen for the IT industry as the initial performance gap was too big between the models. Similar results were found for predicting the EBITDA of manufacturing companies as the choice of words model started to perform better on further predictions. These results conclude that the choice of words model extracted new and valuable information for predicting company metrics. In addition to predicting values new insights are gained in the process. The interpretable factors capture the focus and priorities of the different companies. These insights are easily extracted as this process is automated which could benefit investors and analysts in their research.

However, this approach is not perfect as improvements can still be implemented which will further improve the results and enable it to be implemented on a broader range of industries. As can be seen in the results, the MSE is on the high side for total revenue predictions. This can be improved by scaling all companies and their total revenues to be more comparable. In similar studies this type of scaling is done by taking the ratio of the total revenues with their corresponding value of total assets. This will decrease the variation in the dependent variable because all observations will be compressed into a ratio which relates to the number of revenues a company generates compared to its size. This works well for identical types of industries as businesses are more similar. However, this implementation introduced a significant downwards bias in this research, which resulted in drastically underestimation of these ratios. Being able to implement this standardization and overcoming this bias will decrease the variance in these estimations and change how the factors need to be interpreted. The scale of a company would be accounted for and factors would become more general to all companies of that specific industry.

As seen in the outlier analysis further improvements can be made by implementing a robust estimator in the factor analysis. This can be done by changing the covariance estimator to one that is robust like the MCD estimator. In Python, the implementation of factor analysis is done by means of the minres estimator. Not much documentation exists about this estimator but in the source code it can be found that one could make the factor analysis more robust by substituting the minres estimator with a more robust estimator.

For further research one could investigate different types of information hidden in the text that could be extracted to see if these could further improve the performance. One could also try different items of a yearly report to see if indeed the MD&A item contains most information for these kind of predictions.

At last, when comparing the overall results between the two industries it was found that the model performed significantly worse for the IT industry. As mentioned earlier this could be because the portfolio of companies in the IT industry are to different for the model to comprehend. The IT industry consists of a broad range of totally different companies all operating in different ways. All firm year observations are

treated as separate observations and one could argue that the model does not utilise the fact that each observation is a time point and is related to a specific company. The polynomial fit model does utilize this information. It treats all annual reports of a certain company as a set of observations. One could improve the proposed approach by creating a model which includes this information. Implementing a panel data lasso regression which accounts for the difference in companies and time could improve the model and its predictions. Company specific variables can be added to account for different industry segments or other company specific properties. In addition, this panel data model makes it able to add in past total revenues to further improve these predictions. The choice of words model combined with past total revenues could greatly increase prediction accuracy because more useful information is included in the model and the difference in companies is accounted for.

To conclude, the model has answered the question of whether annual reports contain quantifiable information in the text that can help predictions. In addition to prediction, new additional information is gained in the form of factors which give insights into a company and the state of the market. However, the model still has improvements which need to be made before implementing it in finance. This research proves the fact that textual analysis is a field which can improve financial analysis. Automation, interpretability and extraction of additional information are desired properties in finance and this research demonstrates that these desired properties can be accomplished by means of textual analysis.

References

Agrawal, Shreyash et al. (2018). "Momentum, mean-reversion, and social media: Evidence from stocktwits and twitter". In: *The Journal of Portfolio Management* 44.7, pp. 85–95.

Araci, Dogu (2019). "Finbert: Financial sentiment analysis with pre-trained language models". In: *arXiv preprint arXiv:1908.10063*.

Bartlett, Maurice S (1951). "The effect of standardization on a χ 2 approximation in factor analysis". In: *Biometrika* 38.3/4, pp. 337–344.

Burks, L, M Miller, and R Zadeh (2014). "Rapid estimate of ground shaking intensity by combining simple earthquake characteristics with tweets". In: *10th US Nat. Conf. Earthquake Eng., Front. Earthquake Eng., Anchorage, AK, USA, Jul. 21Y25*.

Campbell, John Y and Samuel B Thompson (2008). "Predicting excess stock returns out of sample: Can anything beat the historical average?" In: *The Review of Financial Studies* 21.4, pp. 1509–1531.

Culotta, Aron, Nirmal Kumar Ravi, and Jennifer Cutler (2016). "Predicting Twitter user demographics using distant supervision from website traffic data". In: *Journal of Artificial Intelligence Research* 55, pp. 389–408.

Das, Sanjiv R and Mike Y Chen (2007). "Yahoo! for Amazon: Sentiment extraction from small talk on the web". In: *Management science* 53.9, pp. 1375–1388.

Devlin, Jacob et al. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805*.

Dziuban, Charles D and Edwin C Shirkey (1974). "When is a correlation matrix appropriate for factor analysis? Some decision rules." In: *Psychological bulletin* 81.6, p. 358.

Eisfeldt, Andrea L and Dimitris Papanikolaou (2014). "The value and ownership of intangible capital". In: *American Economic Review* 104.5, pp. 189–94.

Elliott, W Brooke, Stephanie M Grant, and Frank D Hodge (2018). "Negative news and investor trust: The role of *Firmand#CEOTwitteruse*". In: *Journal of Accounting Research* 56.5, pp. 1483–1519.

Grant, Julia and Larry Parker (2002). "EBITDA!" In: *Research in Accounting Regulation* 15, pp. 205–212.

Grizzle, Gloria A and William Earle Klay (1994). "Forecasting state sales tax revenues: comparing the accuracy of different methods". In: *State & Local Government Review*, pp. 142–152.

Guo, Li, Feng Shi, and Jun Tu (2016). "Textual analysis and machine leaning: Crack unstructured data in finance and accounting". In: *The Journal of Finance and Data Science* 2.3, pp. 153–170.

Gupta, Aaryan et al. (2020). "Comprehensive review of text-mining applications in finance". In: *Financial Innovation* 6.1, pp. 1–25.

Heidari, Maryam and Carsten Felden (2015). "Financial Footnote Analysis: Developing a Text Mining Approach". In.

Hoberg, Gerard and Craig Lewis (2017). "Do fraudulent firms produce abnormal disclosure?" In: *Journal of Corporate Finance* 43, pp. 58–85.

Huang, Norden Eh (2014). *Hilbert-Huang transform and its applications*. Vol. 16. World Scientific.

Jaggi, Mukul et al. (2021). "Text mining of stocktwits data for predicting stock prices". In: *Applied System Innovation* 4.1, p. 13.

Jaseena, KU, Julie M David, et al. (2014). "Issues, challenges, and solutions: big data mining". In: *CS & IT-CSCP* 4.13, pp. 131–140.

Jiang, Ridong, Rafael E Banchs, and Haizhou Li (2016). "Evaluating and combining name entity recognition systems". In: *Proceedings of the Sixth Named Entity Workshop*, pp. 21–27.

Kaalund, SS et al. (2014). "Contrasting changes in DRD1 and DRD2 splice variant expression in schizophrenia and affective disorders, and associations with SNPs in postmortem brain". In: *Molecular psychiatry* 19.12, pp. 1258–1266.

Kaiser, Henry F (1960). "The application of electronic computers to factor analysis". In: *Educational and psychological measurement* 20.1, pp. 141–151.

Kothari, Sabino P, Susan Shu, and Peter D Wysocki (2009). "Do managers withhold bad news?" In: *Journal of Accounting research* 47.1, pp. 241–276.

Lee, BangRae et al. (2018). "About relationship between business text patterns and financial performance in corporate data". In: *Journal of Open Innovation: Technology, Market, and Complexity* 4.1, p. 3.

Lewis, David D et al. (2004). "Rcv1: A new benchmark collection for text categorization research". In: *Journal of machine learning research* 5.Apr, pp. 361–397.

Malo, P. et al. (2014). "Good debt or bad debt: Detecting semantic orientations in economic texts". In: *Journal of the Association for Information Science and Technology* 65.

Mankiw, N Gregory (2021). *Principles of Microeconomics 9e*. Cengage Learning Asia Pte Limited.

McDonald, Barry (2002). "A teaching note on Cook's distance-a guideline". In.

Panchiwala, Shivani and Manan Shah (2020). "A comprehensive study on critical security issues and challenges of the IoT world". In: *Journal of Data, Information and Management* 2.4, pp. 257–278.

Passamani, Giuliana, Roberto Tamborini, and Matteo Tomaselli (2015). "Sustainability vs credibility of fiscal consolidation: A principal components factor analysis for the Euro Zone". In: *The Journal of Risk Finance*.

Pett, Marjorie A, Nancy R Lackey, John J Sullivan, et al. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research*. sage.

Porter, Martin F (2001). "Snowball: A language for stemming algorithms". In.

Price, S McKay et al. (2012). "Earnings conference calls and stock returns: The incremental informativeness of textual tone". In: *Journal of Banking & Finance* 36.4, pp. 992–1011.

Raju, Nambury S et al. (1997). "Methodology review: Estimation of population validity and cross-validity, and the use of equal weights in prediction". In: *Applied Psychological Measurement* 21.4, pp. 291–305.

Rogers, Jonathan L, Douglas J Skinner, and Sarah LC Zechman (2017). "Run EDGAR run: SEC dissemination in a high-frequency world". In: *Journal of Accounting Research* 55.2, pp. 459–505.

Rousseeuw, Peter J (1985). "Multivariate estimation with high breakdown point". In: *Mathematical statistics and applications* 8.283-297, p. 37.

Rousseeuw, Peter J and Katrien Van Driessen (1999). "A fast algorithm for the minimum covariance determinant estimator". In: *Technometrics* 41.3, pp. 212–223.

Sanger, Gary C and John J McConnell (1986). "Stock exchange listings, firm value, and security market efficiency: The impact of NASDAQ". In: *Journal of Financial and Quantitative Analysis* 21.1, pp. 1–25.

Shahi, Amir Mohammad, Biju Issac, and Jashua Rajesh Modapothala (2014). "Automatic analysis of corporate sustainability reports and intelligent scoring". In: *International Journal of Computational Intelligence and Applications* 13.01, p. 1450006.

Sousa, Matheus Gomes et al. (2019). "BERT for stock market sentiment analysis". In: *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, pp. 1597–1601.

Steel, Robert George Douglas, James Hiram Torrie, et al. (1960). "Principles and procedures of statistics." In: *Principles and procedures of statistics.*

Talaviya, Tanha et al. (2020). "Implementation of artificial intelligence in agriculture for optimisation of irrigation and application of pesticides and herbicides". In: *Artificial Intelligence in Agriculture* 4, pp. 58–73.

Tibshirani, Robert (1996). "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288.

Tobias, Sigmund and James E Carlson (1969). "Brief report: Bartlett's test of sphericity and chance findings in factor analysis". In: *Multivariate behavioral research* 4.3, pp. 375–377.

Yang, Yi, Mark Christopher Siy Uy, and Allen Huang (2020). "Finbert: A pretrained language model for financial communications". In: *arXiv preprint arXiv:2006.08097*.

Yin, Ping and Xitao Fan (2001). "Estimating R 2 shrinkage in multiple regression: A comparison of different analytical methods". In: *The Journal of Experimental Education* 69.2, pp. 203–224.

Zhao, Lingyun et al. (2021). "A BERT based sentiment analysis and key entity detection approach for online financial texts". In: pp. 1233–1238.

Appendix

A SIC Codes

The manufacturing SIC codes used are: 2000, 2011, 2013, 2015, 2020, 2024, 2030, 2033, 2040, 2050, 2052, 2060, 2070, 2080, 2082, 2086, 2090, 2092, 2100, 2111, 2200, 2211, 2221, 2250, 2253, 2273, 2300, 2320, 2330, 2340, 2390, 2400.

The IT SIC codes used are: 7370, 7371, 7372, 7373 and 7374.

On the website of the SEC the different categories can be identified. They are not added here because it would increase the length of the document. However, The list is added to the main code. The link to the SEC website: https://www.sec.gov/corpfin/division-of-corporation-finance-standard-industrial-classification-sic-code-list

B Term Frequency-Inverse Document Frequency

The TF-IDF consist of two parts, the term frequency and the inverse document frequency. First the inverse document frequency will be explained and afterwards the term frequency to finally calculate the TF-IDF.

The inverse document frequency (IDF) is a measure to find out how common (or uncommon) a word is amongst the corpus. In other words this measure quantifies the importance of words in a corpus and not in a document. The IDF is calculated in the following equation.

$$idf(t, D) = \log\left(\frac{N}{\text{count}(d \in D : t \in d)}\right) = \log\frac{1+n}{1+df(t)} + 1 \tag{3}$$

Here t is the term (word) and N is the number of documents, d, in corpus D. The denominator is simply the number of documents in which the term, t, appears in. In many programming languages a different formula is used to account for very common words (right part of equation). Here, df(t) is the document frequency of term t which counts the amount of documents the term t occurs in. The reason the IDF is helpful to calculate is because it corrects for words like “of”, “as”, “the”, etc. since they appear frequently in an English corpus. Thus by taking inverse document frequency these common words are accounted for. Minimizing the weight of frequent terms while making infrequent terms have a higher impact.

Now to calculate the TF-IDF the last part needed to be calculated is the term frequency. The term frequency is simply the number of times a word occurs in a document scaled by the amount of words that are present in that document. The TF-IDF is a good way of quantifying the importance of a term in the corpus because it is inversely related to its frequency across documents. The term frequency gives us information on how often a term appears in a document and IDF gives us information about the relative rarity of a term in the collection of documents. By multiplying these values together the TF-IDF is calculated which is stated in the following equation.

$$TF - IDF(t, d, D) = TF(t, d) \cdot IDF(t, D) \tag{4}$$

the higher the TF-IDF score the more important or relevant the term is. Therefore, when a term becomes less relevant, its TF-IDF score will approach 0.

C Bartlett's test of Sphericity and a Kaiser-Meyer-Olkin test

The Bartlett's test of Sphericity is used to test the null hypothesis that the correlation matrix is an identity matrix. An identity correlation matrix means your variables are unrelated and not ideal for factor analysis. A significant statistical test shows that the correlation matrix is indeed not an identity matrix (rejection of the null hypothesis).

$$\chi^2_{p\frac{(p-1)}{2}} = -log(\det|R|)(N-1-\frac{(2p+5)}{6}) \tag{5}$$

Under the null hypothesis that the data is a random sample from the multivariate normal distribution where the covariance matrix is a diagonal matrix, Bartlett showed that this statistic has a chi-square distribution with p(p–1)/2 degrees of freedom (Bartlett 1951).

The Kaiser-Meyer-Olkin (KMO) Test is a measure of how suited your data is for factor analysis. The test measures sampling adequacy for each variable in the model and for the complete model. The statistic is a measure of the proportion of variance among variables that might be common variance. The lower the proportion, the more suited your data is to factor analysis. KMO returns values between 0 and 1. A rule of thumb used for interpreting this statistic is: KMO values between 0.8 and 1 indicate the sampling is adequate, KMO values less than 0.6 indicate the sampling is not adequate and that remedial action should be taken. KMO Values close to zero means that there are large partial correlations compared to the sum of correlations. In other words, there are widespread correlations which are a large problem for factor analysis. The formula for the KMO test is:

$$MO_j = \frac{\sum_{i\neq j} r_{ij}^2}{\sum_{i\neq j} r_{ij}^2 + \sum_{i\neq j} u} \tag{6}$$

where r_ij is the ith,jth element of the correlation matrix and u_ij is the ith,jth element of the partial covariance matrix.

D Validation Criteria

MAE and MSE

In statistics, mean absolute error (MAE) is a measure of errors between paired observations expressing the same phenomenon. Examples of y versus x include comparisons of predicted versus observed, subsequent time versus initial time, and one technique of measurement versus an alternative technique of measurement. MAE is calculated as the sum of absolute errors divided by the sample size.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n} \tag{7}$$

In this equation y_i is the true value and $\hat{y}_i$ is the predicted value. MAE is thus an arithmetic average of the absolute errors $|e_i| = |y_i - \hat{y}_i|$

In statistics, the mean squared error (MSE) of an estimator measures the average of the squares of the errors which in other words is the average squared difference between the estimated values and the actual value. The fact that MSE is almost always strictly positive (and not zero) is because of randomness or because the estimator does not account for information that could produce a more accurate estimate.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 . \tag{8}$$

In this equation y_i is the true value and $\hat{y}_i$ is the predicted value. Taking the root of the MSE will result in the root mean squared error (RMSE)

The R^2 and its variations

In this paper three different variations of the R^2 are used. First, the R^2 and adjusted R^2 will be elaborated. Secondly, the incremental R^2 will be introduced and at last will the pseudo R^2 be explained. The R-Squared, coefficient of determination, can be seen as the proportion of the variation that resides in the dependent variable that is predictable from the "independent" variables. This metric is widely used and often its main purpose is either prediction of future outcomes or the testing of hypothesis. In this thesis its purpose is the former. The R^2 provides a measure for how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model (Steel, Torrie, et al. 1960). The coefficient of determination normally ranges from 0 to 1 whereas a score of 1 is often desired as this means that the model captures the variation of the dependent variable well. The derivation of R^2 is given below.

$$SS_{\mathrm{res}} = \sum_i (y_i - \hat{y}_i)^2 = \sum_i e_i^2 \tag{9}$$

The sum of squares of residuals, also called the residual sum of squares. Where y_i is the true dependent variable, $\hat{y}_i$ is the predicted variable and e_i is the residual.

$$SS_{\mathrm{tot}} = \sum_i (y_i - \bar{y})^2 \tag{10}$$

The total sum of squares (proportional to the variance of the data). Where $\bar{y}$ is the mean of the dependent

variable. This equation is equal to the MSE of a model with only an intercept included (Campbell and Thompson 2008). Combining these equations gives the formula of the coefficient of determination.

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}} = \frac{\sum_i e_i^2}{\sum_i (y_i - \bar{y})^2} \tag{11}$$

The use of an adjusted R^2, $\bar{R}^2$ or R^2_{adj}, is an attempt to account for the phenomenon of the R^2 automatically increasing when extra explanatory variables are added to the model, even when they do not add additional information. In the literature, many different ways of adjusting can be found (Raju et al. 1997). The one used most often to the point that it is typically just referred to as the adjusted R^2, is the correction proposed by Mordecai Ezekiel (Yin and Fan 2001). The definition of the R^2_{adj} is given below.

$$\bar{R}^2 = 1 - \frac{SS_{\text{res}}\,/\text{df}_{\text{res}}}{SS_{\text{tot}}\,/\text{df}_{\text{tot}}} = 1 - \left(1 - R^2\right)\frac{n-1}{n-p} \tag{12}$$

In the left equation, $df_{res} = n - p$ is the degrees of freedom of the estimate of the population variance around the model and $df_{tot}n - 1$ is the degrees of freedom of the estimate of the population variance around the mean. These two terms can be expressed in terms of the sample size n and the number of variables p in the model. Inserting the degrees of freedom and using the definition of R^2, it can be rewritten as the equation on the right. where p is the total number of explanatory variables in the model, and n is the sample size.

The out-of-sample pseudo R^2 is calculated as one minus the ratio of the MSE from a forecasting model to that of a model with only the intercept included (Campbell and Thompson 2008). This benchmark is valid because it captures the predictive power of the historical means of the dependent variables. As was mentioned above this is captured in the calculation of the R^2.

The incremental R^2 is not a widely used definition as lots of different statistics can be used in its place. In a regression or model, the incremental R^2 is the resulting decrease of the R^2 when a variable is removed from such said regression or model. This process is repeated for all the variables in the model such that these values can be compared. This then gives a indication of how much R^2 each variable is adding to the model. In other words, this indicates how much useful information is present in the variable. This is only an indication to compare these variables because when removing a variable, the model is changed and therefore these values can not be used in a other setting then this comparison.

E Manufacturing (Total revenue) - Tables

Table 12: Word lists and top firm-year observations for top 5 factors (1-year ahead) - Manufacturing total revenues

Top 5 generated factors				
Factor 1	Factor 183	Factor 3	Factor 0	Factor 4
Corporate Structure	**Operations**	**Company Governance**	**Finance**	**Beverages**
Top 10 words				
split	abnorm	digit	impair	syrup
insignific	spoilag	medic	tangibl	bottl
divestitur	apb	unalloc	foreign	africa
fewer	normal	singl	pension	concentr
costsw	wast	oversight	analys	case
input	lag	doubl	equit	water
staff	worker	perpetu	method	leadership
coststh	handl	unaudit	definit	beverag
synerg	meal	grew	realize	noncurr
cycl	error	divis	deriv	remeasur
Top 10 firm-year				
MONDELEZ INTERNATIONAL INC - 2007	OMEGA PROTEIN CORP - 2004	PEPSICO INC - 2012	PEPSICO INC - 2015	COCA-COLA CO - 2011
MONDELEZ INTERNATIONAL INC - 2008	OMEGA PROTEIN CORP - 2003	PEPSICO INC - 2014	BUNGE LTD - 2008	COCA-COLA CO - 2010
MEAD JOHNSON NUTRITION CO - 2008	J & J SNACK FOODS CORP - 2015	PEPSICO INC - 2015	PEPSICO INC - 2014	COCA-COLA CO - 2009
0001059581 - 2002	HALLWOOD GROUP INC - 2003	PEPSICO INC - 2011	PEPSICO INC - 2013	COCA-COLA CO - 2015
MONDELEZ INTERNATIONAL INC - 2009	ASHWORTH INC - 2004	PEPSICO INC - 2009	COCA-COLA CO - 2009	COCA-COLA CO - 2013
GUESS INC - 2005	HALLWOOD GROUP INC - 2004	PEPSICO INC - 2010	PEPSICO INC - 2012	COCA-COLA CO - 2012
REYNOLDS AMERICAN INC - 2014	ASHWORTH INC - 2003	PEPSICO INC - 2013	PEPSI BOTTLING GROUP INC - 2007	COCA-COLA CO - 2014
DONNKENNY INC - 2000	B&G FOODS INC - 2004	PEPSICO INC - 2008	COCA-COLA CO - 2010	COCA-COLA CO - 2008
CENTRIC BRANDS INC - 2014	INTERSTATE BAKERIES CORP - 2004	PEPSICO INC - 2007	PEPSI BOTTLING GROUP INC - 2006	COCA-COLA CO - 2007
DONNKENNY INC - 1999	INTERSTATE BAKERIES CORP - 2005	PEPSICO INC - 2004	PEPSICO INC - 2011	COCA-COLA CO - 2006
DONNKENNY INC - 2001	INTERSTATE BAKERIES CORP - 2003	PEPSICO INC - 2006	COCA-COLA CO - 2011	COCA-COLA CO - 2005
UNIFI INC - 2008	TASTY BAKING CO - 2003	PEPSICO INC - 2005	PEPSICO INC - 2010	COCA-COLA CO - 2004
MONDELEZ INTERNATIONAL INC - 2006	HAMPSHIRE GROUP LTD - 2004	PEPSICO INC - 2003	COCA-COLA CO - 2012	COCA-COLA CO - 2003
DONNKENNY INC - 1998	MOHAWK INDUSTRIES INC - 2005	PEPSICO INC - 2002	MOLSON COORS BEVERAGE CO - 2015	COCA-COLA CO - 2002
MEAD JOHNSON NUTRITION CO - 2009	MOHAWK INDUSTRIES INC - 2004	PEPSICO INC - 2001	LEVI STRAUSS & CO - 2002	KEURIG DR PEPPER INC - 2010
FARMER BROTHERS CO - 2003	HERSHEY CO - 2003	0000014693 - 2011	COCA-COLA CO - 2015	PEPSICO INC - 2003
MEAD JOHNSON NUTRITION CO - 2010	OMEGA PROTEIN CORP - 2005	0000014693 - 2012	COCA-COLA CO - 2014	KEURIG DR PEPPER INC - 2009
SANDERSON FARMS INC - 2005	MONTEREY GOURMET FOODS INC - 2003	PEPSICO INC - 1997	COCA-COLA CO - 2013	0001491675 - 2014
SUPERIOR GROUP OF COS INC - 2015	PLANET GREEN HOLDINGS CORP - 2015	0000014693 - 2013	PEPSICO INC - 2009	PEPSICO INC - 2006
FLOWERS FOODS INC - 2012	BUNGE LTD - 2003	PEPSICO INC - 1996	MONDELEZ INTERNATIONAL INC - 2012	KEURIG DR PEPPER INC - 2007

Notes: This table shows the top 5 factors, ranked by their incremental R^2 for the in-sample regressions predicting one-year-ahead total revenues. The factor analysis is run on TF-IDFs which is the term frequency - inverse document frequency. This is done on all available 10-K reports for the manufacturing industry dated from 1995 till 2020. Promax rotation, SMC priors and principal component method is applied. The incremental R^2 for each factor is calculated as to how much the R^2 reduces when the factor is removed from the regression.

Table 13: Word lists and top firm-year observations for top 5 factors (3-year ahead) - Manufacturing total revenues

	Top 5 generated factors			
Factor 1	Factor 18	Factor 3	Factor 0	Factor 4
Beverages	**Fair Trade**	**Supply chain**	**Finance**	**Retail**
Top 10 words				
syrup	align	brazilian	impair	store
africa	obligationsw	farmer	tangibl	occup
leadership	trust	freight	foreign	open
bottl	chain	readil	goodwil	merchandis
case	somewhat	wheat	pension	hand
refer	next	agricultur	analys	traffic
noncurr	stop	crop	equit	websit
concentr	vote	apprec	currenc	markdown
deconsolid	cocoa	oil	method	depart
beverag	durat	farm	deriv	wholesal
Top 10 firm-year				
COCA-COLA CO - 2009	HERSHEY CO - 2010	BUNGE LTD - 2009	BUNGE LTD - 2008	GUESS INC - 2011
COCA-COLA CO - 2010	HERSHEY CO - 2007	BUNGE LTD - 2007	COCA-COLA CO - 2009	LULULEMON ATHLETICA INC - 2011
COCA-COLA CO - 2011	HERSHEY CO - 2006	BUNGE LTD - 2006	PEPSICO INC - 2012	0001085482 - 2007
COCA-COLA CO - 2012	HERSHEY CO - 2011	BUNGE LTD - 2010	PEPSICO INC - 2013	LULULEMON ATHLETICA INC - 2010
COCA-COLA CO - 2013	HERSHEY CO - 2009	BUNGE LTD - 2011	PEPSICO INC - 2011	LULULEMON ATHLETICA INC - 2012
COCA-COLA CO - 2008	HERSHEY CO - 2012	BUNGE LTD - 2005	COCA-COLA CO - 2010	LULULEMON ATHLETICA INC - 2013
COCA-COLA CO - 2007	HERSHEY CO - 2008	BUNGE LTD - 2012	COCA-COLA CO - 2011	GUESS INC - 2010
COCA-COLA CO - 2006	HERSHEY CO - 2005	BUNGE LTD - 2013	PEPSICO INC - 2010	0001085482 - 2008
COCA-COLA CO - 2005	HERSHEY CO - 2004	BUNGE LTD - 2004	PEPSICO INC - 2009	0001085482 - 2006
COCA-COLA CO - 2004	HERSHEY CO - 2003	BUNGE LTD - 2008	COCA-COLA CO - 2012	LULULEMON ATHLETICA INC - 2008
COCA-COLA CO - 2003	HERSHEY CO - 2013	BUNGE LTD - 2003	COCA-COLA CO - 2013	LULULEMON ATHLETICA INC - 2009
COCA-COLA CO - 2002	MONDELEZ INTERNATIONAL INC - 2006	BUNGE LTD - 2002	MOLSON COORS BEVERAGE CO - 2012	0001085482 - 2005
PEPSICO INC - 2003	0001368745 - 2009	ARCHER-DANIELS-MIDLAND CO - 2007	PEPSI BOTTLING GROUP INC - 2005	LULULEMON ATHLETICA INC - 2007
0000014693 - 2012	0000014693 - 2013	ARCHER-DANIELS-MIDLAND CO - 2006	LEVI STRAUSS & CO - 2002	0001085482 - 2004
0001491675 - 2013	LEVI STRAUSS & CO - 2000	SANFILIPPO JOHN B&SON - 2004	OXFORD INDUSTRIES INC - 2012	GUESS INC - 2009
PEPSICO INC - 2008	0001564709 - 2012	ARCHER-DANIELS-MIDLAND CO - 2010	MONDELEZ INTERNATIONAL INC - 2012	BEBE STORES INC - 1999
KEURIG DR PEPPER INC - 2009	MOLSON COORS BEVERAGE CO - 2002	ARCHER-DANIELS-MIDLAND CO - 2008	PEPSICO INC - 2008	"PEETS COFFEE & TEA INC" - 2005
KEURIG DR PEPPER INC - 2010	DIAMOND FOODS INC - 2005	HORIZON ORGANIC HOLDING CORP - 1997	PEPSI BOTTLING GROUP INC - 2004	BEBE STORES INC - 1998
KEURIG DR PEPPER INC - 2007	LEVI STRAUSS & CO - 2002	ARCHER-DANIELS-MIDLAND CO - 2011	OXFORD INDUSTRIES INC - 2013	BEBE STORES INC - 2004
PEPSICO INC - 2005	B&G FOODS INC - 2004	ARCHER-DANIELS-MIDLAND CO - 2012	INTERNATIONAL TEXTLE GRP INC - 2007	BEBE STORES INC - 2001

Notes: This table shows the top 5 factors, ranked by their incremental R^2 for the in-sample regressions predicting three-year-ahead total revenues. The factor analysis is tun on TF-IDFs which is the term frequency - inverse document frequency. This is done on all available 10-K reports for the manufacturing industry dated from 1995 till 2020. Promax rotation, SMC priors and principal component method is applied. The incremental R^2 for each factor is calculated as to how much the R^2 reduces when the factor is removed from the regression.

Table 14: Word lists and top firm-year observations for top 5 factors (5-year ahead) - Manufacturing total revenues

	Top 5 generated factors			
Factor 2	Factor 1	Factor 5	Factor 0	Factor 4
Products	**Commodities**	**Dairy**	**Beverages**	**Expansion**
Top 10 words				
expens	agricultur	chees	digit	africa
inc	brazil	frozen	snack	asia
produc	commod	snack	divis	bottl
year	grain	billion	billion	middl
compan	billion	nut	bottl	billion
chees	nut	commod	middl	digit
grain	asia	milk	commod	snack
livestock	land	middl	compan	commod
turkey	livestock	grain	chees	turkey
pork	divis	asia	africa	brazil
Top 10 firm-year				
LUCILLE FARMS INC - 1998	BUNGE LTD - 2006	MONDELEZ INTERNATIONAL INC - 2004	PEPSICO INC - 2009	COCA-COLA CO - 2004
MONSTER BEVERAGE CORP - 2005	BUNGE LTD - 2009	MONDELEZ INTERNATIONAL INC - 2009	PEPSICO INC - 2004	COCA-COLA CO - 2005
LANCASTER COLONY CORP - 1999	BUNGE LTD - 2005	MONDELEZ INTERNATIONAL INC - 2005	PEPSICO INC - 2007	COCA-COLA CO - 2010
LANCASTER COLONY CORP - 1998	BUNGE LTD - 2007	MONDELEZ INTERNATIONAL INC - 2008	PEPSICO INC - 2008	COCA-COLA CO - 2003
0000742685 - 2000	BUNGE LTD - 2011	MONDELEZ INTERNATIONAL INC - 2003	PEPSICO INC - 2005	COCA-COLA CO - 2009
MONSTER BEVERAGE CORP - 2002	BUNGE LTD - 2010	MONDELEZ INTERNATIONAL INC - 2007	PEPSICO INC - 2006	COCA-COLA CO - 2011
MONSTER BEVERAGE CORP - 2003	BUNGE LTD - 2008	MONDELEZ INTERNATIONAL INC - 2010	PEPSICO INC - 2010	COCA-COLA CO - 2007
MONSTER BEVERAGE CORP - 2004	ARCHER-DANIELS-MIDLAND CO - 2011	MONDELEZ INTERNATIONAL INC - 2006	PEPSICO INC - 2003	COCA-COLA CO - 2008
0001004411 - 2008	BUNGE LTD - 2003	LUCILLE FARMS INC - 1998	PEPSICO INC - 2011	COCA-COLA CO - 2006
0000046640 - 1995	BUNGE LTD - 2004	DEAN FOODS CO - 2003	PEPSICO INC - 2002	ALTRIA GROUP INC - 2011
PULSE BEVERAGE CORP - 2007	BUNGE LTD - 2002	LUCILLE FARMS INC - 1997	PEPSICO INC - 2001	COCA-COLA CO - 2002
0000014693 - 1999	ARCHER-DANIELS-MIDLAND CO - 2007	0000046640 - 2002	PEPSICO INC - 1994	PEPSICO INC - 2006
0000014693 - 2000	ARCHER-DANIELS-MIDLAND CO - 2010	DEAN FOODS CO - 2010	PEPSICO INC - 2000	QUIKSILVER INC - 2009
0000742685 - 2002	ARCHER-DANIELS-MIDLAND CO - 2006	0000046640 - 2004	WARNACO GROUP INC - 2001	INGREDION INC - 2008
MONSTER BEVERAGE CORP - 2011	ARCHER-DANIELS-MIDLAND CO - 2008	CONAGRA BRANDS INC - 2007	PEPSICO INC - 1998	GUESS INC - 2010
0000742685 - 2001	ARCHER-DANIELS-MIDLAND CO - 2009	FLOWERS FOODS INC - 2003	PEPSICO INC - 1995	INGREDION INC - 2009
0000014693 - 1998	SANDERSON FARMS INC - 2010	SANFILIPPO JOHN B&SON - 2011	PEPSICO INC - 1997	INGREDION INC - 2005
0001004411 - 2006	SANDERSON FARMS INC - 2011	J & J SNACK FOODS CORP - 2009	PEPSICO INC - 1996	PEPSICO INC - 2007
0001004411 - 2007	HERSHEY CO - 2006	J & J SNACK FOODS CORP - 2007	PEPSICO INC - 1999	INGREDION INC - 2007
0001347858 - 2007	HERSHEY CO - 2007	0000046640 - 2003	0000768158 - 2001	GUESS INC - 2011

Notes: This table shows the top 5 factors, ranked by their incremental R^2 for the in-sample regressions predicting five-year-ahead total revenues. The factor analysis is run on TF-IDFs which is the term frequency - inverse document frequency. This is done on all available 10-K reports for the manufacturing industry dated from 1995 till 2020. Promax rotation, SMC priors and principal component method is applied. The incremental R^2 for each factor is calculated as to how much the R^2 reduces when the factor is removed from the regression.

F IT (Total revenue) - Tables

Table 15: Word lists and top firm-year observations for top 5 factors (1-year ahead) - IT total revenues

	Top 5 generated factors			
Factor 3	Factor 18	Factor 0	Factor 45	Factor 10
Interactive	**Corporate structure**	**Finance**	**Cloud**	**Service**
Top 10 words				
titl	tender	inc	stabl	holiday
franchis	divestitur	cash	auction	india
publish	consent	expens	issuer	team
launch	default	asset	repurchas	reinvest
interact	repurchas	cost	backlog	sterl
onlin	billion	due	billion	intercompan
video	outsourc	period	discover	billabl
inventor	loan	rate	cloud	client
driven	district	tax	discretionar	outsourc
billion	segment	under	refinanc	softwar
Top 10 firm-year				
ACTIVISION BLIZZARD INC - 2007	AFFILIATED COMPUTER SERVICES - 2005	0001094451 - 2000	ADOBE INC - 2011	VIRTUSA CORP - 2007
ACTIVISION BLIZZARD INC - 2006	AFFILIATED COMPUTER SERVICES - 2006	0001007367 - 1995	ADOBE INC - 2014	VIRTUSA CORP - 2008
ACTIVISION BLIZZARD INC - 2005	0001009575 - 2007	EXTENDED SYSTEMS INC - 1999	OPEN TEXT CORP - 2011	VIRTUSA CORP - 2015
ACTIVISION BLIZZARD INC - 2004	AFFILIATED COMPUTER SERVICES - 2007	PACKETEER INC - 1998	OPEN TEXT CORP - 2015	VIRTUSA CORP - 2009
ELECTRONIC ARTS INC - 2006	AFFILIATED COMPUTER SERVICES - 2004	DIGITAL ORIGIN INC - 1995	ADOBE INC - 2012	VIRTUSA CORP - 2011
ELECTRONIC ARTS INC - 2005	VERISIGN INC - 2009	LHS GROUP INC - 1996	OPEN TEXT CORP - 2009	VIRTUSA CORP - 2012
ACTIVISION BLIZZARD INC - 2015	CA INC - 2006	CORELOGIC INC - 2002	OPEN TEXT CORP - 2012	VIRTUSA CORP - 2013
ACTIVISION BLIZZARD INC - 2012	CA INC - 2004	EBENX INC - 1998	OPEN TEXT CORP - 2010	VIRTUSA CORP - 2010
ACTIVISION BLIZZARD INC - 2011	VERISIGN INC - 2007	ADVENT SOFTWARE INC - 2001	ADOBE INC - 2013	VIRTUSA CORP - 2014
ACTIVISION BLIZZARD INC - 2014	INTERVIDEO INC - 2003	CORELOGIC INC - 2001	VERSANT CORP - 2009	SYNTEL INC - 2006
ELECTRONIC ARTS INC - 2004	FIRST DATA CORP - 1998	ADVENT SOFTWARE INC - 2000	ADOBE INC - 2010	KANBAY INTERNATIONAL INC - 2003
ELECTRONIC ARTS INC - 2010	0000750004 - 2009	CORELOGIC INC - 1994	ADOBE INC - 2015	SYNTEL INC - 2005
ELECTRONIC ARTS INC - 2003	VERISIGN INC - 2008	0001010026 - 1996	0001376321 - 2008	SYNTEL INC - 2014
ACTIVISION BLIZZARD INC - 2013	TRIPADVISOR INC - 2010	EXTENDED SYSTEMS INC - 2000	OPEN TEXT CORP - 2014	IGATE CORP - 2006
ELECTRONIC ARTS INC - 2002	FIRST DATA CORP - 1999	DIGITAL ORIGIN INC - 1996	0001009575 - 2007	SYNTEL INC - 2004
ACTIVISION BLIZZARD INC - 2010	CORELOGIC INC - 2009	CORELOGIC INC - 2003	VERSANT CORP - 2008	SYNTEL INC - 2010
TAKE-TWO INTERACTIVE SFTWR - 2003	CA INC - 2005	CORELOGIC INC - 1995	VERSANT CORP - 2010	SYNTEL INC - 2009
ACTIVISION BLIZZARD INC - 2009	HALIFAX CORP - 2004	CORELOGIC INC - 1996	OPEN TEXT CORP - 2013	INTELLIGROUP INC - 2004
TAKE-TWO INTERACTIVE SFTWR - 2008	0000750004 - 2008	ADVENT SOFTWARE INC - 1999	OPEN TEXT CORP - 2008	IGATE CORP - 2007
TAKE-TWO INTERACTIVE SFTWR - 2000	NOVELL INC - 2007	RSTAR CORP - 2000	CA INC - 2015	INNODATA INC - 2009

Notes: This table shows the top 5 factors, ranked by their incremental R^2 for the in-sample regressions predicting one-year-ahead total revenues. The factor analysis is run on TF-IDFs which is the term frequency - inverse document frequency. This is done on all available 10-K reports for the manufacturing industry dated from 1995 till 2020. Promax rotation, SMC priors and principal component method is applied. The incremental R^2 for each factor is calculated as to how much the R^2 reduces when the factor is removed from the regression.

Table 16: Word lists and top firm-year observations for top 5 factors (3-year ahead forecast) - IT total revenues

	Top 5 generated factors			
Factor 0	Factor 4	Factor 41	Factor 3	Factor 46
Cost and Expenses	**Marketing**	**Restructuring**	**Electronic arts**	**Cloud**
Top 10 words				
inc	content	squar	titl	extraordinar
revenu	page	merger	publish	cloud
includ	media	video	franchis	merger
bas	advertis	restructur	strong	softwar
expens	domain	ventur	inventor	confer
expect	digit	strong	concess	vertic
oper	onlin	microsoft	retail	portal
custom	metric	caption	televis	vice
fin	email	face	onlin	impact
due	softwar	thousand	video	nine
Top 10 firm-year				
0000942650 - 2004	LEAF GROUP LTD - 2011	THQ INC - 2005	ACTIVISION BLIZZARD INC - 2007	ILINC COMMUNICATIONS INC - 2003
0000064463 - 1997	LEAF GROUP LTD - 2012	THQ INC - 2004	ACTIVISION BLIZZARD INC - 2006	VERINT SYSTEMS INC - 2007
0000064463 - 1999	LEAF GROUP LTD - 2010	THQ INC - 2006	ACTIVISION BLIZZARD INC - 2005	PORTAL SOFTWARE INC - 1999
0000942650 - 2003	LEAF GROUP LTD - 2009	ONYX SOFTWARE CORP - 2001	ACTIVISION BLIZZARD INC - 2004	PARK CITY GROUP INC - 2012
0001437925 - 2011	0001142701 - 2009	INTERNAP CORP - 2007	ACTIVISION BLIZZARD INC - 2003	LIVEPERSON INC - 2001
MDSI MOBILE DATA SOLUTIONS - 1997	LOUDEYE CORP - 2000	ELECTRONIC ARTS INC - 2001	ACTIVISION BLIZZARD INC - 2012	0001078404 - 1999
MDSI MOBILE DATA SOLUTIONS - 1999	LOUDEYE CORP - 2001	ARIBA INC - 2003	TAKE-TWO INTERACTIVE SFTWR - 2002	0001372414 - 2013
0000064463 - 2000	LOUDEYE CORP - 1999	RACKSPACE HOSTING INC - 2010	ACTIVISION BLIZZARD INC - 2011	VERTICALNET INC - 2001
CAM COMM SOLUTIONS INC - 2002	LEAF GROUP LTD - 2013	THQ INC - 2003	TAKE-TWO INTERACTIVE SFTWR - 2003	VERSANT CORP - 2002
DATATEC SYSTEMS INC - 1996	0001142701 - 2013	ART TECHNOLOGY GROUP INC - 2003	TAKE-TWO INTERACTIVE SFTWR - 2008	0000750004 - 1994
IDENTIX INC - 1995	0001142701 - 2011	ARIBA INC - 2007	ELECTRONIC ARTS INC - 2004	A D A M INC - 1999
0001376321 - 2006	0001142701 - 2012	SAPIENT CORP - 2001	ACTIVISION BLIZZARD INC - 2013	0000836690 - 2001
IDENTIX INC - 1996	0001142701 - 2005	ART TECHNOLOGY GROUP INC - 2002	ELECTRONIC ARTS INC - 2005	EGAIN CORP - 2012
NETEGRITY INC - 1996	0001142701 - 2010	ARIBA INC - 2004	ACTIVISION BLIZZARD INC - 2002	MCAFEE INC - 2001
NETEGRITY INC - 1997	0001074909 - 2012	ELECTRONIC ARTS INC - 1999	ELECTRONIC ARTS INC - 2003	ULTIMATE SOFTWARE GROUP INC - 2013
0001462223 - 2010	0001142701 - 2006	ARIBA INC - 2005	ACTIVISION BLIZZARD INC - 2010	ADOBE INC - 2007
0001462223 - 2011	SPORTSLINE.COM INC - 1999	ASURE SOFTWARE INC - 2002	TAKE-TWO INTERACTIVE SFTWR - 2005	IPASS INC - 2009
0000064463 - 1998	0001142701 - 2008	INTERNAP CORP - 2011	ELECTRONIC ARTS INC - 2006	NAVTEQ CORP - 2001
0001376321 - 2007	TUCOWS INC - 2012	INTERNAP CORP - 2006	TAKE-TWO INTERACTIVE SFTWR - 2007	0001000495 - 2001
LIVERAMP HOLDINGS INC - 2005	TUCOWS INC - 2010	VERITAS SOFTWARE CORP - 1999	TAKE-TWO INTERACTIVE SFTWR - 2000	ONSTREAM MEDIA CORP - 2011

Notes: This table shows the top 5 factors, ranked by their incremental R^2 for the in-sample regressions predicting three-year-ahead total revenues. The factor analysis is tun on TF-IDFs which is the term frequency - inverse document frequency. This is done on all available 10-K reports for the manufacturing industry dated from 1995 till 2020. Promax rotation, SMC priors and principal component method is applied. The incremental R^2 for each factor is calculated as to how much the R^2 reduces when the factor is removed from the regression.

Table 17: Word lists and top firm-year observations for top 5 factors (5-year ahead forecast) - IT total revenues

	Top 5 generated factors			
Factor 19	Factor 0	Factor 2	Factor 10	Factor 49
Strategy	**Expenses**	**Business model**	**Revenue**	**Operations**
Top 10 words				
restructur	inc	failur	fsp	erp
exit	includ	serious	sfas	sg&a
peopl	bas	manufactur	noncontrol	billabl
sfas	cash	telecommun	sab	local
microsoft	expens	distributor	hedg	noncontrol
distributor	revenu	local	emea	microsoft
ventur	recogn	san	construct	segment
hedg	account	merger	index	translat
stabl	asset	prefer	revenu	construct
digit	due	microsoft	video	outsourc
Top 10 firm-year				
SAPIENT CORP - 2001	PEOPLESOFT INC - 1995	0001142701 - 2001	ELECTRONIC ARTS INC - 2008	CMTSU LIQUIDATION INC - 2006
ENTRUST INC - 2002	MDSI MOBILE DATA SOLUTIONS - 1997	MICROMUSE INC - 1997	NEXTGEN HEALTHCARE INC - 2008	AMERICAN SOFTWARE -CL A - 2010
0001094451 - 2000	IDENTIX INC - 1995	0001142701 - 2003	0001376321 - 2011	AMERICAN SOFTWARE -CL A - 2011
SAPIENT CORP - 2003	IDENTIX INC - 1996	0001087955 - 1998	ANSYS INC - 2007	CMTSU LIQUIDATION INC - 2008
SAPIENT CORP - 2002	LIVERAMP HOLDINGS INC - 2005	MICROMUSE INC - 1998	SOLERA HOLDINGS INC - 2009	EDGEWATER TECHNOLOGY INC - 2011
0001142701 - 2010	NETEGRITY INC - 1996	0001142701 - 2002	ASPEN TECHNOLOGY INC - 2008	EDGEWATER TECHNOLOGY INC - 2010
BROADVISION INC - 2002	NETEGRITY INC - 1997	0001087955 - 1999	ZIX CORP - 2007	AMERICAN SOFTWARE -CL A - 2009
BROADVISION INC - 2001	0001040596 - 2004	0001010026 - 1996	VERINT SYSTEMS INC - 2007	CMTSU LIQUIDATION INC - 2009
0001142701 - 2011	0001040596 - 2000	LOUDEYE CORP - 1999	FALCONSTOR SOFTWARE INC - 2007	EDGEWATER TECHNOLOGY INC - 2009
EARTHLINK HOLDINGS CORP - 2002	0001376321 - 2006	0001023731 - 2004	SOLERA HOLDINGS INC - 2008	CMTSU LIQUIDATION INC - 2005
BROADVISION INC - 2003	EPICEDGE INC - 1994	I2 TECHNOLOGIES INC - 1999	0001009575 - 2007	IGATE CORP - 2001
SAPIENT CORP - 2004	GTECH HOLDINGS CORP - 1999	0001087955 - 2000	0001040896 - 2007	0000910638 - 2005
ART TECHNOLOGY GROUP INC - 2003	0000815185 - 1997	I2 TECHNOLOGIES INC - 2000	MICROS SYSTEMS INC - 2008	EDGEWATER TECHNOLOGY INC - 2008
NETMANAGE INC - 2001	0000815185 - 1996	SPORTSLINE.COM INC - 1997	VERINT SYSTEMS INC - 2008	CMTSU LIQUIDATION INC - 2007
BORLAND SOFTWARE CORP - 2001	0001462223 - 2010	CACI INTL INC -CL A - 2002	ACI WORLDWIDE INC - 2007	IGATE CORP - 2002
CADENCE DESIGN SYSTEMS INC - 2004	0001376321 - 2007	ARIBA INC - 2003	0001376321 - 2010	0001272550 - 2011
0001009575 - 2001	0000915016 - 1995	CACI INTL INC -CL A - 2003	ASPEN TECHNOLOGY INC - 2007	CMTSU LIQUIDATION INC - 2004
BMC SOFTWARE INC - 2005	0001462223 - 2011	ASCENTIAL SOFTWARE CORP - 1996	ELECTRONIC ARTS INC - 2007	AMERICAN SOFTWARE -CL A - 2008
INTERNAP CORP - 2001	0001040596 - 2001	0001063085 - 1998	0001438231 - 2007	0000910638 - 2006
SAPIENT CORP - 2005	0001040596 - 2002	CACI INTL INC -CL A - 2004	OPEN TEXT CORP - 2008	CMTSU LIQUIDATION INC - 2003

Notes: This table shows the top 5 factors, ranked by their incremental R^2 for the in-sample regressions predicting five-year-ahead total revenues. The factor analysis is run on TF-IDFs which is the term frequency - inverse document frequency. This is done on all available 10-K reports for the manufacturing industry dated from 1995 till 2020. Promax rotation, SMC priors and principal component method is applied. The incremental R^2 for each factor is calculated as to how much the R^2 reduces when the factor is removed from the regression.

G Manufacturing (EBITDA) - Tables

Table 18: Word lists and top firm-year observations for top 5 factors (1-year ahead forecast) - Manufacturing EBITDA

Top 5 generated factors				
Factor 3	Factor 2	Factor 1	Factor 0	Factor 42
Sales	**Disposition**	**Legal**	**Risks**	**Innovation**
Top 10 words				
gross	divestitur	impos	tangibl	headquart
sale	chees	prohibit	pension	recours
increas	favor	enforc	alloc	stabil
new	mix	regul	retir	innov
indemnif	unfavor	submit	contribut	around
obsolesc	billion	track	exposur	premium
bad	snack	rule	currenc	popul
fiscal	resolut	user	jurisdict	fift
caution	saturday	enact	exclud	cautionar
profit	africa	legisl	hedg	spend
Top 10 firm-year				
REYNOLDS AMERICAN INC - 2009	MONDELEZ INTERNATIONAL INC - 2004	ALTRIA GROUP INC - 2011	BUNGE LTD - 2008	UNDER ARMOUR INC - 2011
AMER ITALIAN PASTA CO -CL A - 2004	MONDELEZ INTERNATIONAL INC - 2009	ALTRIA GROUP INC - 2012	PEPSI BOTTLING GROUP INC - 2007	UNDER ARMOUR INC - 2010
MONDELEZ INTERNATIONAL INC - 2012	MONDELEZ INTERNATIONAL INC - 2005	ALTRIA GROUP INC - 2015	PEPSI BOTTLING GROUP INC - 2006	UNDER ARMOUR INC - 2012
REYNOLDS AMERICAN INC - 2010	MONDELEZ INTERNATIONAL INC - 2008	ALTRIA GROUP INC - 2013	LEVI STRAUSS & CO - 2002	UST INC - 2005
REYNOLDS AMERICAN INC - 2008	MONDELEZ INTERNATIONAL INC - 2010	ALTRIA GROUP INC - 2014	OXFORD INDUSTRIES INC - 2012	SPECTRUM ORGANIC PRODUCTS - 2002
SMITHFIELD FOODS INC - 2007	MONDELEZ INTERNATIONAL INC - 2007	REYNOLDS AMERICAN INC - 2014	PEPSI BOTTLING GROUP INC - 2005	GYMBOREE CORP - 2004
UNIFI INC - 2006	MONDELEZ INTERNATIONAL INC - 2003	REYNOLDS AMERICAN INC - 2013	OXFORD INDUSTRIES INC - 2013	VOLCOM INC - 2006
UNIFI INC - 2005	MONDELEZ INTERNATIONAL INC - 2006	REYNOLDS AMERICAN INC - 2012	PEPSI BOTTLING GROUP INC - 2004	HAGGAR CORP - 2002
SMITHFIELD FOODS INC - 2008	MONDELEZ INTERNATIONAL INC - 2012	REYNOLDS AMERICAN INC - 2011	OXFORD INDUSTRIES INC - 2014	VOLCOM INC - 2005
REYNOLDS AMERICAN INC - 2004	MONDELEZ INTERNATIONAL INC - 2011	REYNOLDS AMERICAN INC - 2010	MOLSON COORS BEVERAGE CO - 2012	UNDER ARMOUR INC - 2015
MONDELEZ INTERNATIONAL INC - 2013	MONDELEZ INTERNATIONAL INC - 2014	REYNOLDS AMERICAN INC - 2008	INTERNATIONAL TEXTLE GRP INC - 2007	INTERNATIONAL TEXTLE GRP INC - 2010
HALLWOOD GROUP INC - 2003	MONDELEZ INTERNATIONAL INC - 2015	REYNOLDS AMERICAN INC - 2009	MOLSON COORS BEVERAGE CO - 2015	UNDER ARMOUR INC - 2013
MEAD JOHNSON NUTRITION CO - 2007	MONDELEZ INTERNATIONAL INC - 2013	0000844856 - 2015	OXFORD INDUSTRIES INC - 2011	UNDER ARMOUR INC - 2014
MEAD JOHNSON NUTRITION CO - 2008	0000046640 - 2002	REYNOLDS AMERICAN INC - 2007	MONDELEZ INTERNATIONAL INC - 2012	INTERNATIONAL TEXTLE GRP INC - 2009
PEPSI BOTTLING GROUP INC - 2003	0000046640 - 2004	0000059440 - 2015	MOLSON COORS BEVERAGE CO - 2014	COCA COLA CONSOLIDATED INC - 2004
COLUMBIA SPORTSWEAR CO - 2002	0000046640 - 2005	REYNOLDS AMERICAN INC - 2006	PEPSI BOTTLING GROUP INC - 2003	PINNACLE FOODS INC - 2012
UNDER ARMOUR INC - 2011	GENERAL MILLS INC - 2015	0000059440 - 2014	MOLSON COORS BEVERAGE CO - 2011	BUNGE LTD - 2008
RICEBRAN TECHNOLOGIES - 2007	GALAXY NUTRITIONAL FOODS INC - 2003	REYNOLDS AMERICAN INC - 2005	OXFORD INDUSTRIES INC - 2015	0000789073 - 2011
REYNOLDS AMERICAN INC - 2006	GENERAL MILLS INC - 2009	0000059440 - 2010	MOLSON COORS BEVERAGE CO - 2013	CAMPBELL SOUP CO - 2010
UNIFI INC - 2002	0000046640 - 2003	0000059440 - 2013	OXFORD INDUSTRIES INC - 2008	VF CORP - 2012

Notes: This table shows the top 5 factors, ranked by their incremental R^2 for the in-sample regressions predicting one-year-ahead total revenues. The factor analysis is tun on TF-IDFs which is the term frequency - inverse document frequency. This is done on all available 10-K reports for the manufacturing industry dated from 1995 till 2020. Promax rotation, SMC priors and principal component method is applied. The incremental R^2 for each factor is calculated as to how much the R^2 reduces when the factor is removed from the regression.

Table 19: Word lists and top firm-year observations for top 5 factors (3-year ahead forecast) - Manufacturing EBITDA

	Top 5 generated factors			
Factor 1	Factor 0	Factor 176	Factor 36	Factor 11
Investors	**Impairment**	**Peformance**	**Partnerships**	**Export**
Top 10 words				
prime	impair	compan	tast	unobserv
libor	discount	assetsin	page	observ
bear	exceed	quantit	arrear	hierarch
loan	calcul	limit	read	input
equal	tangibl	yen	california	put
waiver	assess	likelihood	distributor	life
violat	foreign	relationship	shown	core
arrear	return	self	territor	guidanc
minus	analys	defens	exclus	export
instal	pension	lag	eurodollar	element
Top 10 firm-year				
GALAXY NUTRITIONAL FOODS INC - 2001	LEVI STRAUSS & CO - 2002	KATE SPADE & CO - 2012	FLOWERS FOODS INC - 2012	CULP INC - 2008
CHAUS BERNARD INC - 2001	OXFORD INDUSTRIES INC - 2012	INTERNATIONAL TEXTLE GRP INC - 1995	FLOWERS FOODS INC - 2011	LANDEC CORP - 2012
CHAUS BERNARD INC - 2002	OXFORD INDUSTRIES INC - 2013	TREEHOUSE FOODS INC - 2013	0001491675 - 2009	LANDEC CORP - 2011
GALEY & LORD INC - 1999	PEPSI BOTTLING GROUP INC - 2004	B&G FOODS INC - 2009	FLOWERS FOODS INC - 2010	FRESHPET INC - 2013
0000844161 - 2011	PEPSI BOTTLING GROUP INC - 2005	FARMER BROTHERS CO - 2011	DARLING INGREDIENTS INC - 2006	FLOWERS FOODS INC - 2009
CHAUS BERNARD INC - 2003	INTERNATIONAL TEXTLE GRP INC - 2007	KATE SPADE & CO - 2011	DARLING INGREDIENTS INC - 2005	LANDEC CORP - 2013
GALEY & LORD INC - 1998	OXFORD INDUSTRIES INC - 2011	FARMER BROTHERS CO - 2012	RALCORP HOLDINGS INC - 2006	LANDEC CORP - 2010
CHAUS BERNARD INC - 2006	PEPSI BOTTLING GROUP INC - 2003	0000041017 - 1997	RALCORP HOLDINGS INC - 2005	CRAFT BREW ALLIANCE INC - 2007
CHAUS BERNARD INC - 2007	OXFORD INDUSTRIES INC - 2008	INTERNATIONAL TEXTLE GRP INC - 1996	0001579823 - 2013	0000042136 - 2010
CHAUS BERNARD INC - 2004	OXFORD INDUSTRIES INC - 2010	HAMPSHIRE GROUP LTD - 2010	0001564709 - 2013	0000042136 - 2009
"PILGRIMS PRIDE CORP" - 2008	OXFORD INDUSTRIES INC - 2009	0001315257 - 2004	0000100979 - 1996	0000042136 - 2011
GALAXY NUTRITIONAL FOODS INC - 2000	0000046640 - 2012	INTERNATIONAL TEXTLE GRP INC - 2008	0000838875 - 2012	0000042136 - 2012
0000844161 - 2013	0001287151 - 2008	GUESS INC - 2011	LORILLARD INC - 2007	0001066923 - 2013
CHAUS BERNARD INC - 2005	CONAGRA BRANDS INC - 2013	COCA COLA CONSOLIDATED INC - 2012	MONTEREY GOURMET FOODS INC - 2001	0001066923 - 2012
INTERNATIONAL TEXTLE GRP INC - 1995	HERSHEY CO - 2011	KEURIG DR PEPPER INC - 2010	0000838875 - 2013	0001287151 - 2007
RICEBRAN TECHNOLOGIES - 2007	RALPH LAUREN CORP - 2012	RALPH LAUREN CORP - 2011	0001564709 - 2012	"PEETS COFFEE & TEA INC" - 2007
INTERNATIONAL TEXTLE GRP INC - 2004	RALPH LAUREN CORP - 2008	0000866873 - 2002	BOSTON BEER INC -CL A - 2006	LANDEC CORP - 2009
INVENTURE FOODS INC - 2001	KATE SPADE & CO - 2005	DEAN FOODS CO - 2011	LORILLARD INC - 2009	DEAN FOODS CO - 2007
CROWN CRAFTS INC - 2012	RALPH LAUREN CORP - 2013	B&G FOODS INC - 2008	KATE SPADE & CO - 2005	GUESS INC - 2011
0001287151 - 2008	HERSHEY CO - 2012	OXFORD INDUSTRIES INC - 2010	MONTEREY GOURMET FOODS INC - 2000	0001287151 - 2009

Notes: This table shows the top 5 factors, ranked by their incremental R^2 for the in-sample regressions predicting three-year-ahead total revenues. The factor analysis is run on TF-IDFs which is the term frequency - inverse document frequency. This is done on all available 10-K reports for the manufacturing industry dated from 1995 till 2020. Promax rotation, SMC priors and principal component method is applied. The incremental R^2 for each factor is calculated as to how much the R^2 reduces when the factor is removed from the regression.

Table 20: Word lists and top firm-year observations for top 5 factors (5-year ahead forecast) - Manufacturing EBITDA

	Top 5 generated factors			
Factor 23	Factor 0	Factor 1	Factor 40	Factor 12
Seasonal	**Finance**	**Legislation**	**Suppliers**	**Beverages**
Top 10 words				
outerwear	impair	valley	obsolet	belief
spring	discount	legisl	sole	cooper
footwear	exceed	appeal	sampl	bottl
wear	partial	embed	sport	sponsor
columbia	sheet	master	doubt	aluminum
sportswear	goodwil	florida	thereb	followsth
access	tangibl	litig	apparel	handl
geograph	calcul	court	supplier	withdraw
sport	foreign	commenc	excess	water
season	analys	disput	commiss	perpetu
Top 10 firm-year				
COLUMBIA SPORTSWEAR CO - 2006	LEVI STRAUSS & CO - 2002	0000059440 - 2005	SPORT-HALEY HOLDINGS INC - 2003	COCA COLA CONSOLIDATED INC - 2009
COLUMBIA SPORTSWEAR CO - 2009	OXFORD INDUSTRIES INC - 2011	0000059440 - 2006	SPORT-HALEY HOLDINGS INC - 2004	COCA COLA CONSOLIDATED INC - 2010
COLUMBIA SPORTSWEAR CO - 2007	PEPSI BOTTLING GROUP INC - 2003	0000059440 - 2007	SPORT-HALEY HOLDINGS INC - 2002	COCA COLA CONSOLIDATED INC - 2007
COLUMBIA SPORTSWEAR CO - 2004	MOLSON COORS BEVERAGE CO - 2011	0000059440 - 2004	WARNACO GROUP INC - 2003	COCA COLA CONSOLIDATED INC - 2011
COLUMBIA SPORTSWEAR CO - 2005	INTERNATIONAL TEXTLE GRP INC - 2007	0000059440 - 2008	SPORT-HALEY HOLDINGS INC - 2005	COCA COLA CONSOLIDATED INC - 2006
COLUMBIA SPORTSWEAR CO - 2008	OXFORD INDUSTRIES INC - 2010	0000059440 - 2003	SPORT-HALEY HOLDINGS INC - 2001	COCA COLA CONSOLIDATED INC - 2005
COLUMBIA SPORTSWEAR CO - 2003	OXFORD INDUSTRIES INC - 2008	0000059440 - 2010	WARNACO GROUP INC - 2004	COCA COLA CONSOLIDATED INC - 2004
COLUMBIA SPORTSWEAR CO - 2010	0001287151 - 2008	0000059440 - 2011	WARNACO GROUP INC - 2002	COCA COLA CONSOLIDATED INC - 2002
COLUMBIA SPORTSWEAR CO - 2011	OXFORD INDUSTRIES INC - 2009	0000059440 - 2009	0001311538 - 2011	COCA COLA CONSOLIDATED INC - 2001
COLUMBIA SPORTSWEAR CO - 2002	KATE SPADE & CO - 2008	0000059440 - 2002	SPORT-HALEY HOLDINGS INC - 2000	PEPSI BOTTLING GROUP INC - 2003
COLUMBIA SPORTSWEAR CO - 2001	INTERNATIONAL TEXTLE GRP INC - 2008	0000059440 - 2001	0001311538 - 2008	0001315257 - 2004
COLUMBIA SPORTSWEAR CO - 2000	KATE SPADE & CO - 2009	LOUISIANA-PACIFIC CORP - 1998	0000771298 - 2000	0001491675 - 2011
G-III APPAREL GROUP LTD - 2011	RALPH LAUREN CORP - 2008	0001315257 - 2004	0001311538 - 2009	0000909987 - 2001
G-III APPAREL GROUP LTD - 2009	KATE SPADE & CO - 2005	ALTRIA GROUP INC - 2011	CENTRIC BRANDS INC - 2011	PVH CORP - 2011
G-III APPAREL GROUP LTD - 2010	HERSHEY CO - 2011	LOUISIANA-PACIFIC CORP - 2000	0001311538 - 2010	"PILGRIMS PRIDE CORP" - 2001
JONES GROUP INC - 2002	0000866873 - 1997	"PILGRIMS PRIDE CORP" - 2008	0000838875 - 2008	COCA COLA CONSOLIDATED INC - 1999
0001085482 - 2003	VF CORP - 2010	0000864264 - 2000	QUIKSILVER INC - 2002	HAIN CELESTIAL GROUP INC - 2009
JONES GROUP INC - 2003	RALPH LAUREN CORP - 2011	0000838875 - 2008	JONES SODA CO - 2009	NATIONAL BEVERAGE CORP - 2009
WARNACO GROUP INC - 2004	MOLSON COORS BEVERAGE CO - 2005	0001347858 - 2011	CENTRIC BRANDS INC - 2010	MOLSON COORS BEVERAGE CO - 2003
WARNACO GROUP INC - 2005	VF CORP - 2009	RALPH LAUREN CORP - 2004	PULSE BEVERAGE CORP - 2007	COCA COLA CONSOLIDATED INC - 1995

Notes: This table shows the top 5 factors, ranked by their incremental R^2 for the in-sample regressions predicting five-year-ahead total revenues. The factor analysis is run on TF-IDFs which is the term frequency - inverse document frequency. This is done on all available 10-K reports for the manufacturing industry dated from 1995 till 2020. Promax rotation, SMC priors and principal component method is applied. The incremental R^2 for each factor is calculated as to how much the R^2 reduces when the factor is removed from the regression.