Transformer-based Summarization and Sentiment Analysis of SEC 10-K Annual Reports for Company Performance Prediction

Hsin-Ting Hsieh
HWR Berlin, Germany
hthsieh.tw@gmail.com

Diana Hristova
HWR Berlin, Germany
diana.hristova@hwr-berlin.de

Abstract

Annual reports published by companies contain important insights regarding their performance and are often analyzed in a manual, subjective manner. We address this point by combining the streams of research on text summarization and topic modelling with the one on sentiment analysis. Our approach consists of the steps of text summarization using BERTSUMEXT, topic modelling with LDA, sentiment analysis with FinBERT, and performance prediction with Decision Trees and Random Forest. The result provides decision makers with an interpretable and condensed representation of the content of annual reports, together with its relationship to future company performance. We evaluate our approach on 10-K reports, demonstrating both its interpretability for analysts and explanatory power regarding future company performance.

1. Introduction

In the past years, the volume of available information from the financial sector has increased tremendously [1] with companies producing every year a massive number of reports. These reports are known as financial disclosures and convey company business situations in numerical and textual ways. Among them, annual reports are considered being one of the most representative sources of information [2]. They disclose companies' operating and financial activities over the past year and are commonly used by credit analysts, accountants and investors to evaluate financial performance and make investment decisions. Traditionally, the focus would lie solely on the analysis of backward-looking quantitative financial metrics as a basis for making decisions. However, as many studies have shown [3, 4, 5, 6, 7, 8], annual reports contain one additional type of extremely valuable information related to the company's future performance, namely qualitative textual information. As opposed to financial metrics, this information contains forward-looking statements on *topics*, such as risk factors, industry outlook or M&A [5, 7, 8]. These statements could have a positive or negative *sentiment* depending on the company's expectations [4, 8, 9]. Thus analyzing the textual information, in addition to the backward-looking financial metrics, provides a more thorough picture of the company and leads to better decisions.

However, annual reports are mostly reviewed manually, in a time-consuming, subjective and complex process. The resulting assessment could differ among analysts [5] and even be inconsistent for the same person and different companies. Furthermore, research has shown that both the length and redundant words in annual reports have increased over time, resulting in more review time and information overload [4]. The language complexity of the reports has increased as well, making their manual review, especially by smaller investors and on time to support investment decisions, almost impossible [4]. The above issues can be addressed by an automated review process. This was recognized by other researchers, who applied different natural language processing (NLP) techniques to retrieve the *topics* and text *sentiment* in the reports.

Topics are derived using unsupervised learning, with the application of topic modelling and summarization techniques [10, 11, 12]. Topic modelling in terms of Latent Dirichlet Allocation (LDA) aims at assigning each text to a distribution of a set of topics, while text summarization produces a condensed and informative summary of a long document. For instance, Dyer et al. [12] apply LDA to capture the topic evolution of 10-K annual reports and explain the increase in the length of annual reports over time. Also, Zheng et al. [13] show that state-of-the-art NLP methods can generate high-quality summaries of annual reports.

Sentiment is the focus of the field of sentiment analysis, which examines its relationship with financial performance indicators such as stock price, future earnings [4] or period returns [9] using supervised learning. Sentiment is calculated based on the structured representation of the unstructured texts. This can be derived either by a bag-of-words approach (frequency-based) or an embeddings approach (continuous-vector-space-based) for both words or sentences [14].

URI: https://hdl.handle.net/10125/79550
978-0-9981331-5-7

The literature on sentiment analysis demonstrates that sentiment provides valuable information regarding company performance [8, 9, 15]. It should thus be considered in addition to financial metrics in decision making. However, existing works do not focus on the interpretability of the model results for decision makers. We define interpretability as the ability of humans to understand model results and refer the interested reader to the XAI literature [16]. In particular, for annual reports of realistic length, neither bag-of-words nor embeddings can be interpreted by analysts. This is crucial, because management, regulators and the society expect that when applying complex NLP techniques, a human holds the ultimate responsibility for the taken decision [17]. Due to this accountability, even for high performing models, analysts would still like to understand the model results (i.e. them to be interpretable) to trust them [18]. If not, they would conduct a manual review instead, thus making the model useless. The same holds for regulators and the society, who would allow the use of such approaches, only if they can be interpreted [19].

The literature on summarization and topic modelling addresses this point by generating short and interpretable results. However, it does not examine their relationship with the company performance. As a result, the report still has to be additionally manually reviewed (even though less than before), leading to the above issues of time, subjectivity and complexity. Thus, there is a research gap represented by our research question:

RQ: *How can we automatically extract the topics and sentiment from annual reports in a condensed and interpretable way and use the result to support investment decisions?*

In this work, we aim answering this question by proposing a state-of-the-art NLP methodology that generates the summaries of annual reports together with their topics and sentiment and relates them to company performance in an interpretable way. We use the summaries and not the whole text to reduce noise in the topics and thus provide better results. Based on previous works, we focus on analyzing the U.S. Securities and Exchange Commission (SEC) 10-K annual reports.

Our contribution consists of combining the two existing streams of research in an analysis pipeline, which generates a high performing and interpretable model result. After applying this pipeline to a given annual report, the analyst is provided with a quick and objective prediction of the company performance. It can be used for decision support and also in combination with financial metrics. Additionally, to facilitate trust in this prediction, our approach generates the summaries together with their topics and sentiment, making it interpretable. Our hypothesis is as follows:

H1: *The generated model results are interpretable and have high explanatory power in terms of future company performance*

The paper is structured as follows: in the next section, we discuss in detail the literature in the above two research fields and derive the existing research gap. Then we present our methodology in section 3, followed by its evaluation in section 4. Finally, in section 5 we derive main conclusions and paths for future research.

2. Related work and background

2.1. Text summarization and topic modelling

The application of text summarization, is very promising in the financial domain [20]. Recently, a study revealed that the automatically produced summary of earnings releases helps investors reduce positive bias and leads to a more conservative value evaluation than the manager-generated summary [10]. The authors used several approaches to conduct text summarization. Among them, LexRank [21], an algorithm for obtaining the relative importance of sentences, performed the best. However, the 2020 Financial Narrative Processing Workshops used LexRank as one of the baselines for the Financial Narrative Summarization shared task [22] and most solutions outperformed LexRank. The task participants were asked to automatically summarize UK financial annual reports. A wide spectrum of methodologies was used, ranging from rule-based methods to deep learning models. Among all models, SUMSUM [13] achieved the highest performance based on Rouge-2 F1 score.

SUMSUM uses a BERT-based classifier to classify whether a section should be included in the summary. BERT is a transformer-based state-of-the-art NLP model for obtaining embeddings. Zheng et al. [13] first derive the BERT-embeddings for each section and then add a linear layer to obtain a classification output. Compared to other methodologies, such as the pointer network and bidirectional long short-term memory, SUMSUM demonstrates that the pretrained BERT model with further adjustment can already achieve impressive results. However, because the data and the models from Zheng et al. [13] are not publicly available, SUMSUM cannot be used for this study. Therefore, BERTSUMEXT [11] is considered as an alternative BERT-based model for text summarization.

BERTSUMEXT was pre-trained using news data from the CNN/DailyMail dataset, which contains news articles and associated highlights. The final output layer of BERTSUMEXT is a classifier which helps the model obtain the importance score for each sentence. The model ranks these sentences by their scores, and selects the top-3 sentences as the summary. BERTSUMEXT

achieved state-of-the-art performance on various datasets, but no study has applied it to annual reports. We propose this in section 3. We would like to note here that such an application comes with limitations stemming from the training dataset. It differs from our dataset both in terms of type (news vs. reports) and domain (general vs. financial). Thus, the model may be performing worse than on the training dataset. In section 5, we discuss this limitation and its solutions.

Another technique, similar to text summarization, is LDA topic modelling. It assumes that every document can be represented as a probability distribution over a set of topics, where each topic is a probability distribution over a set of words. Therefore, the representation is interpretable and shorter than the initial report. This approach was applied successfully in the literature, for instance to analyze the evolution of topics in annual reports [12] or to compare the distribution of topics between earnings conference calls and the subsequent analyst reports [23]. To sum up, both text summarization and topic modelling provide condensed and interpretable representation of texts. However, they do not focus on its relationship with company performance. Sentiment analysis address this point.

2.2. Sentiment analysis in finance

The field of sentiment analysis in finance has a high significance and long history. Initially, the dictionary-based approach was widely used, which classifies the sentiment of words using a predefined dictionary [24]. Afterwards, machine learning (ML) algorithms were applied to analyze company reports, news, or Twitter data and predict important performance indicators in a supervised manner. For instance, Pagolu et al. [25] used Twitter data with financial information and trained a Random Forest (RF) model to detect the sentiment regarding the mentioned financial entity.

With the development of deep learning techniques in NLP, different studies have employed various neural architectures for financial sentiment analysis [26]. For instance, Kraus and Feuerriegel [26] applied a long short-term memory neural network on ad hoc announcements to predict stock market movement. However, training such models requires a vast volume of labeled data, which is not realistic in the financial domain [27]. Therefore, fine-tuning pretrained models, such as FinBERT, has become a promising solution. FinBERT uses BERT's architecture but further pretrains on TRC2-financial, a financial text corpus consisting of 1.8 M news articles. The model is additionally fine-tuned for sentiment classification using the Financial PhraseBank dataset, which has 4,845 English sentences from financial news, and a continuous sentiment score from the FiQA Sentiment dataset. FinBERT achieved higher accuracy than ULMFit and ELMo [27], which are two other pre-trained language models. However, no research has applied it to annual reports before. Due to different data types, here a similar limitation as with BERTSUMEXT exists and is discussed in section 5.

Most existing works retrieve data from news or tweets and focus on analyzing the sentiment of those sources and its effect on company stock price. When it comes to annual reports, several studies were conducted for 10-K reports based on bag-of-words models (e.g. [3, 4] using RF) or word-level embeddings [28, 29, 30], all considering single words in isolation. However, as most finance keywords are context-sensitive, according to Liu et al. [30], such word-based approaches are limited. To extend the understanding of report texts on a sentence level, Du et al. [14] introduce a sentence-level risk-labeled dataset to retrieve risk financial sentiment phrases and further use them on risk classification. As most previous studies using 10-K reports [24, 31], Du et al. [14] also only analyzes Item 7 of the 10-K report.

To sum up, existing works in the field of sentiment analysis focus on the relationship between text representation and performance indicators. Thus, they lack interpretability for decision makers, as most annual reports consist of several hundred pages. To address this gap, we propose an approach that combines the fields of text summarization/topic modelling and sentiment analysis to derive an output that is both interpretable and generates insights on company performance. For this, we first apply BERTSUMEXT as a state-of-the-art summarization technique. In the second and third steps, running in parallel, we derive the sentiment and topics of the summaries, using FinBERT and LDA, respectively. This leads to a condensed and interpretable representation of the reports. To determine its implications for company performance, in the fourth step, we estimate a ML model for predicting future stock price growth. We focus on Decision Trees (DT) and RF, due to their interpretability and common use in the literature. In the next section, we present our approach.

3. Methodology

Figure 1 shows the four steps of our methodology, which are presented in detail the following subsections.

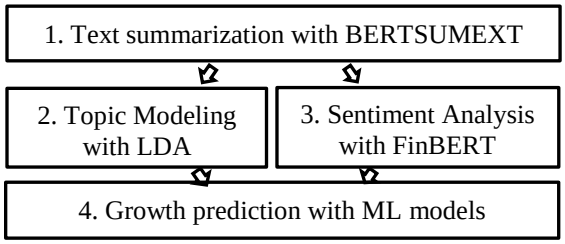

Figure 1. Methodology overview

3.1. Text summarization with BERTSUMEXT

In Step 1, we apply BERTSUMEXT to reduce the amount of text in 10-K annual reports. Since BERT and thus BERTSUMEXT can process only a maximum sequence of 512 tokens (corresponding to at least 510 words), we extract the summaries by first splitting the reports into their items (i.e. sections) and then splitting the items into chunks of 512 tokens. Afterwards, we determine the summary of each chunk and concatenate the results into item summaries. The output of this step is the condensed content of each item.

3.2. Topic Modeling with LDA

With the result from Step 1, decision makers are provided with an output that is much shorter than the initial annual report. However, it may still be too long and difficult to interpret. Therefore, to condense and structure the result further, LDA is applied to the entire summary corpus. We use the summaries and not to the whole texts, as they focus on the important parts and therefore noise in the topics is reduced. Normally, each summary would be assigned to the topic with the highest probability. However, here, we follow a second-best approach to prevent assigning topics that are too specific for a particular company. For instance, in the annual report of Coca Cola, a topic containing words like "cola", "bottle", "bottler" could be the most probable. However, it could be irrelevant for all other companies, leading to the loss of this data point. Thus, if the most probable topic belongs to the summaries of only one company, it is replaced by the topic with the second-highest probability. Therefore, topics containing too specific company information are eliminated. LDA requires a fixed number of topics as input parameter which we determine using the topic coherence score [32]. This score evaluates the average of the semantic similarity between words in a topic. The higher the value, the higher the semantic similarity.

The output of this step is a topic for each summary. Thus, it represents the summary information in a condensed and interpretable form. If the topic distribution would be considered instead, then both interpretability and model performance would suffer due to the number of features and sparsity. Still, the topics do not provide the sentiment of the summaries, which is important when analyzing company performance. This is done in the next step, which can run in parallel with this step, since it uses the same input.

3.3. Sentiment analysis with FinBERT

As mentioned above, we use FinBERT, a BERT-based model, which was pre-trained and fine-tuned for sentiment analysis in the financial domain. FinBERT processes input text at a sentence level, so that each sentence obtains a sentiment category and a sentiment score in the range [-1, 1], with 1 being the most positive. Since Step 1 generates summaries on an item level, the sentence-level sentiment must be aggregated before further analysis. Therefore, three aggregated variables are calculated: *sentiment-score-full*, *sentiment-score* and *sentiment-score-strong*. *Sentiment-score-full* is derived by taking the mean of all sentence sentiment scores. However, this approach may lead to mean values close to zero for most documents, since it is expected that most sentences have neutral sentiment (see 4.4). Therefore, we additionally calculate *sentiment-score* as the mean of the scores that belong only to positive or negative categories, based on FinBERT's prediction. Additionally, we determine *sentiment-score-strong* as the mean of the sentiment values for sentences of the more common positive or negative category. If more sentences have a negative sentiment than a positive one, then *sentiment-score-strong* is the mean of the scores of only the negative sentences and vice versa.

After this step, the item summary with the summary topic and sentiment are available and can be used by the decision maker as an interpretable output to quickly review a report from different perspectives. Still, it is not possible to state how and how well this output reflects company performance. This is done in the next step.

3.4. Growth prediction with ML models

In order to examine the relationship between the generated interpretable output and company performance, we estimate two tree-based models (DT and RF) to predict the future stock price growth. They were chosen because of their use in the sentiment analysis literature and interpretable results.

DT [33] consist of nodes and branches, where the nodes represent the variables and the branches stand for certain decision rules on the variable values. The final nodes in the tree are called leaves and determine the value of the target variable. DT are interpretable and efficient and as opposed to other interpretable models capable of capturing non-linear relationships. However, they also have a risk of overfitting, especially with more complex relationships. To address the latter, restrictions can be set on the tree growth, such as the minimum number of samples in a leaf node. Alternatively, RF [34] can be applied, which is an ensemble comprising numerous individual DT. As such, it has better generalizability and performance. However, due to the high number of trees, it is also less interpretable. Thus, there is a trade-off between the performance and interpretability for DT and RF. Still they are appropriate for this task as they can model complex relationships

better than linear models and have more interpretable results than deep learning models.

We evaluate the performance of the tree-based models using the mean squared error (MSE) metric. The reason for this choice is the continuous target variable (future stock price growth rate) and the fact that MSE is the most commonly-used evaluation metric for regression problems [35]. It is the average squared difference between the predicted and actual value. Therefore, a lower MSE value indicates a more accurate prediction, zero being a perfect one.

The output from this final step is a ML model that predicts the future stock price growth rate based on the summaries' topics and sentiment from the previous steps. This prediction can be used to judge the future company performance and thus make better decisions. Additionally, due to the interpretability of the results, the decision maker can understand the reasons for the prediction, making its use more probable. In the next section, we evaluate our approach.

4. Evaluation

In this section, we first describe the data used for the evaluation, followed by the results in each of the four steps. Additional results can be found here: https://github.com/hsiehkl/Summarization-and-Sentiment-Analysis-of-SEC-10K.

4.1. Data Description

4.1.1. SEC 10-K reports. The primary data comprise the textual content of 10-K reports. The form 10-K has a strict structure which contains 15 items and each item requires companies to disclose corresponding information. After examining the items and also based on the literature, we chose to use the following six items as they are considered important in the literature and consist of textual parts rather than tables: Item 1, Item 1A, Item 3, Item 7, Item 7A, Item 8. We focus on 20 companies in the Dow Jones Industrial Average (DJIA) Index and extract the reports having a filling date between 31.12.2008 and 03.01.2021 as in that period the structure of the reports remained stable. The 20 companies are across 12 industries from Information Technology, FMCG, Aerospace and Defense, etc. and in total 236 reports were extracted.

During text preprocessing, each report is split into several item documents. However, some of the documents were removed since they seemed to be written with a template. For example, IBM often uses a sentence like *"Refer to note M, 'Contingencies and Commitments,' on pages 119 to 121 of IBM's 2017 Annual Report to Stockholders, which is incorporated herein by reference"* in Item 3 and only modifies the page number and the year for different years. Such texts are considered uninformative, because they do not convey the company development over the years and were excluded. After preprocessing, Item 1 and Item 1A had mostly remained and Item 3 and Item 8 were removed more frequently. Finally, 685 items remained in the dataset. Detailed description of our dataset can be found on GitHub under: 'Dataset Statistics.pdf'.

4.1.2. Stock price. Step 4 of our methodology uses the summaries' topics and sentiment to predict company performance. Instead of directly predicting the annual growth rate, we use the adjusted growth rate (company growth rate minus the growth rate of the DJIA Index) as the target variable. Subtracting the growth of the index from the company growth can offset the overall market trend and focus on the premium of individual companies. We use the fiscal year end date of the report to predict the growth for the year starting at that date. The stock price data are obtained from Yahoo Finance.

This completes the description of the data. In the next subsections, we present the results of the evaluation of our methodology.

4.2. Text summarization with BERTSUMEXT

Step 1 consists of extracting summaries using BERTSUMEXT, which are then additionally structured to generate further interpretable outputs in the next steps. No gold standard exists to quantify the quality of the results, so the evaluation focuses on randomly choosing some data points to check the overall performance. The sentences selected by the model are colored in red in the original input text extracts in Figures 2, 3 and 4, which show both high- and low-quality summaries. The whole text is on GitHub in 'Summarization Examples.pdf'. Figure 2 indicates that Johnson & Johnson had just closed a trial with Guidant in 2015 and recorded a gain in this case.

> In recent years, Johnson & Johnson has received numerous requests from a variety of United States Congressional Committees to produce information relevant to ongoing congressional inquiries. It is the policy of Johnson & Johnson […] a merger agreement between Johnson & Johnson and Guidant. In June 2011, Guidant filed a motion for summary judgment and in July 2014, the judge denied Guidant's motion. The trial concluded in January 2015 and in February 2015 […] Johnson & Johnson dismissed its action against Guidant with prejudice. The Company recorded a gain associated with this transaction in fiscal first quarter of 2015. In June 2009, following the public announcement that Ortho-Clinical Diagnostics, Inc. (OCD) had received a grand jury subpoena […]

Figure 2. Summarization result extract (1)

Additionally, in Figure 3, the first two highlighted sentences point out that the tax expense significantly

affected 3M's financial results, but the third sentence is not very informative. These two examples show that BERTSUMXT can well recognize sentences disclosing company's situation and important events, which readers can further explore, if considered relevant.

> During the fourth quarter of 2017, 3M recorded a net tax expense of $762 million related to the enactment of the Tax Cuts and Jobs Act (TCJA). The expense is primarily related […] the Company also provides non-GAAP measures that adjust for the net impact of enactment of the TCJA. This item represents a significant charge that impacted the Company's financial results. Income, earnings per share, and the effective tax rate are all measures for which 3M provides the reported GAAP measure and an adjusted measure. The adjusted measures are not in accordance with, nor are they a substitute for, GAAP measures. The Company considers these non-GAAP measures in evaluating and managing the Company's […]

Figure 3. Summarization result extract (2)

In contrast to such good summaries, Figure 4 depicts an issue with BERTSUMEXT tending to select the first three or the last sentences as a summary. This is because the model was trained with news data, and typically, the most informative sentences in the news are at the beginning or the end of the article.

> The extent of 3M's operations involves dealing with uncertainties and judgments in the application of complex tax regulations in a multitude of jurisdictions. The final taxes paid are dependent upon many factors, including negotiations with taxing authorities in various jurisdictions and resolution of disputes arising from federal, state, and international tax audits. The Company recognizes potential liabilities and records tax liabilities for anticipated tax audit issues in the United States and other tax jurisdictions based on its estimate of whether, and the extent to which, additional taxes will be due. The Company follows guidance provided by ASC 740, Income Taxes […]

Figure 4. Summarization result extract (3)

Overall, this approach helps reduce the number of words in a given report by around 88%, from an average of 46,802 to 5,866 words. Thus, the reader is provided with an already condensed and interpretable version of the report. However, it still may be too long and also it is not clear how it influences company performance. To address those, we apply Steps 2, 3 and 4 below.

4.3. Topic Modeling with LDA

By applying LDA, we derive hidden topics from the summaries, thus facilitating their interpretability and further analysis. We set the number of topics to 20 by analyzing the development of the coherence score for topics varying from 10 to 40 with a step of 5. We used the *genism* library with default parameters. To evaluate the results, the number of documents assigned to each topic was analyzed and can be found on GitHub in 'Document Topic Distribution.png'. We see that Topic 1 was assigned to more than 100 documents, followed by Topic 13 with around 80 documents. All topics can be found on GitHub under 'Twenty Topics from LDA.pdf'.

Topic 1 is characterized by the following top 10 most probable words:

'0.022"product"+0.020"result"+0.015"affect"+ 0.012"operation"+0.011"market"+0.011"cost"+0.011" include"+0.010"financial"+0.010"increase"+0.010"cu stomer"'.

Topic 1 addresses general company characteristics in terms of the financial and operational results regarding the products, markets, and customers, which are an important part of every annual report. Thus, it is natural that this is the most frequent topic. Also, few topics capture company business content. For example, Topic 4 has the following most probable words:

'0.031"service"+0.021"customer"+0.020"product" +0.015"technology"+0.015"network"+0.013"market" +0.011"include"+0.009"provide"+0.008"solution"+0. 008"datum"'.

Topic 4 is assigned to the summaries of Intel, Cisco, IBM, and Verizon. These companies all provide various technology solutions as products or services to customers. Similarly, Topic 15 is represented by the following most probable words:

'0.033"card"+0.013"merchant"+0.012"capital"+ 0.012"risk"+0.011"service"+0.011"payment"+0.011"i nclude"+0.010"financial"+0.009"credit"+0.009"comp any"'.

The companies with summaries about Topic 15 are The Travelers Companies, American Express Company and J.P Morgan, which are all financial services corporations. This implies that a topic can contain the information of a certain industry.

Another topic that attracts attention is Topic 0. Two eye-catching keywords here are "beverage" and "bottle". Among the companies in the data, only Coca-Cola has business concerning these two words. Therefore, it is expected that the LDA model would assign this topic only to documents from Coca-Cola, and with the second-best approach proposed in the methodology to no summary at all. After investigating this issue, Topic 0 was the most probable one for both Coca-Cola and American Express, due to data quality issues. Thus, Topic 0, which contains strong company information, remains in the dataset.

Apart from the topics discussed above, some topics contain words such as "tax", "risk", "insurance", "interest", "loan" and "debt". For example, Topic 16 has the following most probable words:

'0.078"rate"+0.062"risk"+0.045"interest"+0.039" market"+0.029"price"+0.026"currency"+0.025"instru ment"+0.023"debt"+0.021"investment"+0.021"fix"'.

The keywords "risk", "debt", "interest" and "investment" found together may indicate that

companies are aware of exchange rate and interest rate risks regarding their debt or investment. The fluctuation of these rates can cause huge impact on companies' financial performance.

We can see that the results from this step are easily interpretable and thus facilitate the quick and objective analysis of the annual report by the reader. In order to additionally determine whether the summaries are positive or negative, in the next subsection we proceed with Step 3, sentiment analysis.

4.4. Sentiment analysis with FinBERT

In this step, we apply FinBERT to each summary. As the result is a sentiment (category and score) for each sentence, we additionally derive the aggregated sentiment variables as discussed in section 3.3. To better interpret the results, in Figures 5 and 6, we present two word clouds for both the sentences with a positive sentiment category and the ones with a negative one. In the positive word cloud words such as "higher", "revenue", "customer" and "growth" are the most frequent ones. On the other hand, the words "operation", "cost", "decrease" and "impact" are seen in the negative cloud. The common words like "increase", "compared", "market" and "sale" demonstrate that the model captures the sentiment beyond separate words and considers the context around them.

Figure 5. Positive word cloud

Figure 6. Negative word cloud

The sentiment analysis results are further evaluated on both sentence and item level. For sentence level, because it is impossible to review all sentences, a few documents are randomly chosen to investigate the model performance. The example in Table 1 is based on Item 7 of the Boeing 10-K report for 2014.

Table 1. Sentiment analysis result (1)

Neutral	While our principal operations are in the U.S., we conduct operations in many countries and rely on an extensive network of international partners, key suppliers and subcontractors.
Positive	Together with strong demand growth, we expect lower oil prices will improve airline profitability in 2015.
Negative	Changes in our forecasts or decreases in the value of our common stock could cause book values of certain operations to exceed their fair values which may result in goodwill impairment charges in future periods.

As mentioned above, most sentences (74.52%) are classified as neutral and the text of the neutral example in Table 1 is an objective description of Boeing's operation facts. The positive text indicates that the company expected a more profitable market due to strong demand and lower material costs in the next year, which might be a good sign that investors would want to know. Finally, the negative text conveys a possible goodwill impairment in the future, which may adversely affect the company. After considering the sentiment of sentences, Table 2 shows the average values for the three numerical aggregated sentiment variables per item. In all cases, Item 1 and Item 7 have positive average values, while the values for Item 1A, Item 3, Item 7A and Item 8 are negative. The result of Item 1A, Risk Factors is in line with previous research [36] indicating that this item mainly contains negative sentiment.

Table 2. Average aggregated sentiment scores

	Sentiment-score-full	Sentiment-score	Sentiment-score-strong
Item 1	0.12	0.42	0.52
Item 1A	-0.20	-0.53	-0.78
Item 3	-0.36	-0.15	-0.40
Item 7	0.04	0.06	0.22
Item 7A	-0.01	-0.12	-0.13
Item 8	-0.01	-0.26	-0.55

The sentiment analysis produced by the model adds additional forward-looking information and can be used in the next step together with the topics to predict company performance.

4.5. Growth prediction with DT and RF

In this step we estimate two tree-based ML models (DT and RF), to derive the relationship between the

topics and summary sentiment on the one side and the future stock price growth on the other side. We always use one aggregated sentiment variable at a time to avoid multicollinearity. The results are evaluated using MSE and 80%/20% train/test split. Additionally, we apply Grid Search with 10-fold cross validation to determine the best model hyperparameters. The final models all have a minimum sample leaf of six and 400 trees in the RF. We then applied these models to the test set to derive the performance and avoid information leakage.

Table 3 demonstrates the performance of all models. Residual distribution of the best models can be found in 'Evaluation of Tree Models.pdf'. It reveals that all MSE values are very small. This is due to the range of the adjusted growth rate between -1 and 1. Moreover, little differentiation exists in the results across the three sentiment score types. Still, the MSE values of DT are higher than the ones for RF. The following subsections review the feature interpretation of the two best models for DT and RF.

Table 3. MSE (test set) for different models

	Sentiment-score-full	Sentiment-score	Sentiment-score-strong
DT	0.03607	0.03491	0.03668
RF	0.03001	0.03057	0.02961

4.5.1. Evaluation of DT. For DT, the best model uses *sentiment-score*. Therefore, we investigate this tree in terms of its feature importance (Figure 7) and tree structure (Figure 8). In Figure 7, *sentiment-score* is the most important feature, followed by *Topic 15* and then by *Item 7*. Also, we see that after *Topic 10* all topics and items have a zero importance, demonstrating that few features dominated the growth of the tree.

For a visual understanding of the model, the top right part of the tree is shown in Figure 8. The splitting rule in the root node is based on the value of the sentiment score. If the score is lower than -0.881, then a terminal leaf is reached, and the returned prediction value equals -0.059. The MSE value of this leaf is 0.005, which is very small. This result demonstrates that if the sentiment of a summary is extremely negative regardless of the topic or item, the expected company performance is also negative with a high probability. However, if the sentiment score is greater than -0.881, the data move to the decision node that checks whether the topic is 15. Topic 15, as discussed above, is related to financial and capital services and risk. If following the path of Topic 15 (right), and if the sentiment score is greater than -0.514, the results of most leaves imply positive future growth with a low MSE. In contrast, if the sentiment score is less than -0.514, the tree surprisingly still returns a positive prediction result, and its value is even higher than data with a more positive sentiment score. However, the MSE of this leaf is relatively high, reaching 0.13, and it has only seven samples. With the model's minimum sample leaf set to six, this node cannot be further split to obtain a more precise result. One possible solution for this issue may be providing more data in future with more samples on this decision path. In general, the DT indicates that the sentiment and topics extracted from the summary can predict company performance in an interpretable way.

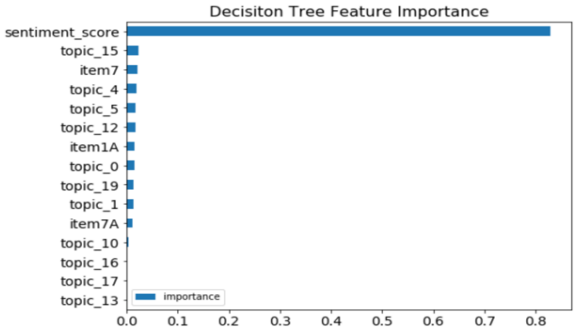

Figure 7. Feature importance (DT)

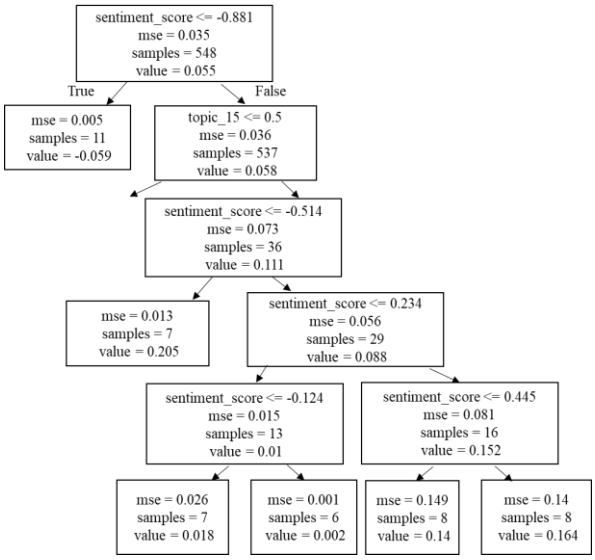

Figure 8. Top Right of the DT

4.5.1. Evaluation of RF. The RF model using *sentiment-score-strong* has the lowest MSE value (see Table 3), so we discuss this model here. As shown in Figure 9, like DT, the most and second-most important features are *sentiment-score-strong* and *Topic 15*, but followed by *Topic 1* instead of *Item 7*. Compared to DT, RF has more features with positive importance for performance prediction.

The model consists of 400 trees. Therefore, instead of plotting all trees, the evaluation was conducted using the Python *treeinterpreter* package, which decomposes each prediction into its bias (training set mean) plus the sum of each feature contribution. Three examples are

shown in Table 4 to examine the effect of the sentiment score on the final predicted value.

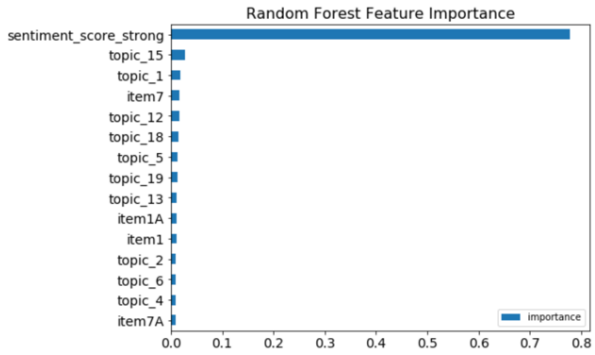

Figure 9. Feature importance (RF)

Table 4. Prediction decomposition

	Sentiment Score Strong	Sentiment Contribution	Predicted Value	True Value
Example 1	0.05929	0.04327	0.0794	0.0736
Example 2	-0.76	-0.0638	-0.0190	-0.0401
Example 3	-0.7979	0.0623	0.1135	0.1141

Examples 1 and 2 demonstrate that a positive sentiment score usually causes a positive sentiment contribution, resulting in a positive predicted value. In contrast, a negative score may result in a negative sentiment contribution and lead to a negative predicted value (Example 2). However, there are exceptions such as Example 3, with a strong negative sentiment score and a positive sentiment contribution. The data point is assigned to Topic 1, and it belongs to Item 1A. This may imply that if a report discusses company financial and operational results regarding products, markets, and customers in the Risk Factors item, it might not be a negative signal, even if it has a negative sentiment.

The result of this section shows that our methodology not only generates a condensed and interpretable output, but also how to use it to predict future company performance with high explanatory power (H1). As opposed to existing methods in the field of sentiment analysis, we can provide the reasons for a given company performance in terms of the text sentiment and important topics (cf. Figure 8 and Table 4). As opposed to methods of topic modelling and summarization, we can make a prediction regarding this performance. Finally, as opposed to works only focusing on financial metrics, we consider forward-looking information (cf. Topic 16).

5. Conclusion

In this paper, we answer RQ1 by presenting a methodology for the extraction of a condensed and interpretable output from the textual parts of annual reports as well as its relationship to company performance. As a result, decision makers can process the lengthy reports quickly and objectively, saving manual effort. Moreover, they know and understand the implications of a report's content for future company performance. Our methodology combines the two streams of research of text summarization/topic modelling and sentiment analysis using state-of-the-art NLP methods. It consists of four steps: 1) creating summaries of the reports with BERTSUMEXT 2) deriving the summaries' topics with LDA, 3) summaries' sentiment analysis with FinBERT and 4) a tree-based ML model for the prediction of future stock price growth. The above steps can be implemented as a pipeline in an IS used for the automated analysis of annual reports in addition to financial metrics. This IS would be developed together with financial experts, who could also enhance the topic interpretability.

We evaluate our approach on a dataset consisting of 10-K annual reports extracted based on the companies in the DJIA. Our results show that we can successfully provide analysts with an automated, efficient and objective review of annual reports. It removes lengthy report content by 88% and further represents it as interpretable topics concerning general company characteristics, industry-specific factors and risk factors. Also, it derives its sentiment, helping decision makers better relate the output to the company's situation. We determine this relation in Step 4, where both DT and RF demonstrate that sentiment is the most important feature. It is followed by topics on general company characteristics, industry-specific factors and Item 7, known for its important role in the literature.

Our approach also has some limitations. In particular, we have small amount of labeled data for both summarization and sentiment analysis. The transformer-based models address this point, but they are both trained on news datasets. Thus, the second limitation is the type of data used for model pretraining. Both BERTSUMEXT and FinBERT are not trained on annual reports. Here, news from a general domain and financial news are analyzed, respectively. Thus, the summarization model training data differs in its type and domain from the application data. For the sentiment analysis, only the type differs (news vs. reports). This can be solved by training the models on annual reports. However, for summarization, this is a purely manual task, because there are no summaries for annual reports available. Similarly, for sentiment, it requires generating labelled data. Alternatively, unsupervised summarization models could be examined. Finally, future work could also consider predicting the next-day stock price after the date a report is published instead of the annual growth.

6. References

[1] A. Gupta, V. Dengre, H. A. Kheruwala, and M. Shah, "Comprehensive review of text-mining applications in finance," *Financial Innovation*, vol. 6, no. 1. Springer Science and Business Media Deutschland GmbH, Dec. 01, 2020.

[2] C. Masson and P. Paroubek, "NLP Analytics in Finance with DoRe: a French 257M Tokens Corpus of Corporate Annual Reports," 2020.

[3] D. Hristova, J. Probst, and E. Eckrich, "RatingBot: A Text Mining Based Rating Approach," 2018.

[4] T. Kang, D.-H. Park, and I. Han, "Beyond the numbers: The effect of 10-K tone on firms' performance predictions using text analytics," *Telemat. Informatics*, vol. 35, pp. 370–381, 2018.

[5] M. Butler and V. Kešelj, "Financial Forecasting Using Character N-Gram Analysis and Readability Scores of Annual Reports," in *Advances in Artificial Intelligence*, 2009, pp. 39–51.

[6] B. Miller, "The Effects of Reporting Complexity on Small and Large Investor Trading," *Account. Rev.*, vol. 85, pp. 2107–2143, 2010.

[7] H. You and X. jun Zhang, "Financial reporting complexity and investor underreaction to 10-k information," *Rev. Account. Stud.*, 2009

[8] Y. Qian and Y. Sun, "The Correlation Between Annual Reports' Narratives and Business Performance: A Retrospective Analysis," *SAGE*, vol. 11, no. 3, 2021

[9] T. Loughran and B. Mcdonald, "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks," *J. Finance*, vol. 66, no. 1, pp. 35–65, Feb. 2011

[10] E. Cardinaels, S. Hollander, and B. J. White, "Automatic summarization of earnings releases: attributes and effects on investors' judgments," *Rev. Account. Stud.*, vol. 24, no. 3, pp. 860–890, Sep. 2019

[11] Y. Liu and M. Lapata, "Text summarization with pretrained encoders," in *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2020, pp. 3730–3740.

[12] T. Dyer, M. Lang, and L. Stice-Lawrence, "The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation," *J. Account. Econ.*, vol. 64, no. 2, pp. 221–245, 2017

[13] S. Zheng, A. Lu, and C. Cardie, "SUMSUM@FNS-2020 Shared Task," COLING, 2020.

[14] C. H. Du, M. F. Tsai, and C. J. Wang, "Beyond Word-level to Sentence-level Sentiment Analysis for Financial Reports," 2019.

[15] F. Li, "The information content of forward- looking statements in corporate filings-A naïve bayesian machine learning approach," *J. Account. Res.*, vol. 48, no. 5, pp. 1049–1102, 2010

[16] C. Molnar, *Interpretable Machine Learning*. 2019.

[17] OECD, "What are the OECD Principles on AI?," 2019. https://www.oecd.org/going-digital/ai/principles/

[18] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[19] E. Toreini *et al.*, "Technologies for Trustworthy Machine Learning: A Survey in a Socio-Technical Context." 2021.

[20] M. El-Haj, P. Rayson, M. Walker, S. Young, and V. Simak, "In search of meaning: Lessons, resources and next steps for computational analysis of financial discourse," *J. Bus. Financ. Account.*, 2019.

[21] G. Erkan and D. R. Radev, "LexRank: Graph-based lexical centrality as salience in text summarization," *J. Artif. Intell. Res.*, 2004

[22] M. El-Haj, A. Abura'ed, M. Litvak, N. Pittaras, and G. Giannakopoulos, "The Financial Narrative Summarisation Shared Task (FNS 2020)," *aclweb.org*, pp. 1–12, 2020

[23] A. H. Huang, R. Lehavy, A. Y. Zang, and R. Zheng, "Analyst information discovery and interpretation roles: A topic modeling approach," in *Management Science*, Jun. 2018, vol. 64, no. 6, pp. 2833–2855.

[24] S. Taylor and V. Keselj, "Using Extractive Lexicon-based Sentiment Analysis to Enhance Understanding ofthe Impact of Non-{GAAP} Measures in Financial Reporting," in *Proceedings of the Second Workshop on Financial Technology and Natural Language Processing*, 2020, pp. 40–46.

[25] V. S. Pagolu, K. N. Reddy, G. Panda, and B. Majhi, "Sentiment analysis of Twitter data for predicting stock market movements," 2017.

[26] M. Kraus and S. Feuerriegel, "Decision support from financial disclosures with deep neural networks and transfer learning," *Decis. Support Syst.*, 2017

[27] D. T. Araci, "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models," *arXiv*. 2019.

[28] C. Nopp and A. Hanbury, "Detecting risks in the banking system by sentiment analysis," 2015.

[29] M. F. Tsai and C. J. Wang, "On the risk prediction and analysis of soft information in finance reports," *Eur. J. Oper. Res.*, 2017

[30] Y. W. Liu, L. C. Liu, C. J. Wang, and M. F. Tsai, "FIN10K: A web-based information system for financial report analysis and visualization," 2016.

[31] A. K. Davis and I. Tama-Sweet, "Managers' Use of Language Across Alternative Disclosure Outlets: Earnings Press Releases versus MD&A," *Contemp. Account. Res.*, 2012

[32] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," 2015.

[33] J. R. Quinlan, "Induction of Decision Trees," *Mach. Learn.*, 1986

[34] L. Breiman, "Random forests," *Mach. Learn.*, 2001

[35] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013.

[36] M. Azimi and A. Agrawal, "Is the Sentiment in Corporate Annual Reports Informative? Evidence from Deep Learning," *SSRN Electron. J.*, Oct. 2018