Learning to Compare Financial Reports for Financial Forecasting

Ross Koval[1,3], **Nicholas Andrews**[2], **and Xifeng Yan**[1]

[1]University of California, Santa Barbara
[2]Johns Hopkins University
[3]AJO Vista
rkoval@ucsb.edu

Abstract

Public companies in the US are required to publish annual reports that detail their recent financial performance, present the current state of ongoing business operations, and discuss future prospects. However, they typically contain over 25,000 words across all sections, large amounts of industry and legal jargon, and a high percentage of boilerplate content that does not change much year-to-year. These unique characteristics present challenges for many generic pretrained language models because it is likely that only a small percentage of the long report reflects salient information that contains meaningful signal about the future prospects of the company. In this work, we curate a large-scale dataset of paired financial reports and introduce two novel, challenging tasks of predicting long-horizon company risk and correlation that evaluate the ability of the model to recognize cross-document relationships with complex, nuanced signals. We explore and present a comprehensive set of methods and experiments, and establish strong baselines designed to learn to identify subtle similarities and differences between long documents. Furthermore, we demonstrate that it is possible to predict company risk and correlation solely from the text of their financial reports and further that modeling the cross-document interactions at a fine-grained level provides significant benefit. Finally, we probe the best performing model through quantitative and qualitative interpretability methods to reveal some insight into the underlying task signal.

1 Introduction

Investors are faced with the consumption of a myriad of textual datasets relevant to financial markets, spanning genres such as news, social media posts, and financial reports. Public companies in the US are required to publish annual reports detailing the current operations of the firm, recent financial performance, and discussing future prospects. However, these reports contain over 25,000 words in length and large amounts of financial and legal jargon. As noted in Cohen et al. (2020), this length and linguistic complexity have increased significantly over time as a result of increased government regulations and business complexity, making it difficult for investors to efficiently process the salient information contained in these reports.

Annual Report - 2014	Annual Report - 2015
Our operations and facilities are subject to extensive federal, state and local laws and regulations relating to the exploration for, and the development, production and transportation of, oil and natural gas, and operating safety...	Our operations and facilities are subject to extensive federal, state and local laws and regulations relating to the exploration for, and the development, production and transportation of, oil and natural gas, and operating safety...
Results of Operations	Depending upon future prices and our production volumes, our cash flows from our operating activities may not be sufficient to fund our capital expenditures, and we may need additional borrowings. ...If commodity prices remain low, we may also recognize further impairments of our producing oil and gas properties if the expected future cash flows from these properties becomes insufficient to recover their carrying value, and we may recognize additional impairments.
Our oil and gas sales increased $35.5 million (9%) in 2013 to $420.3 million from $384.8 million in 2012. Oil sales in 2013 increased by $50.7 million (28%) from 2012 while our natural gas sales decreased by $15.2 million (8%) from 2012. The increase in oil sales was attributable to the 29% growth in oil production offset by a 1% decrease in our realized oil prices in 2013...	**Results of Operations**

Figure 1: Comparison of a sample of passages from consecutive annual reports from the validation dataset of the Risk Prediction task that highlights the salient sentences that were added that potentially indicate an increase in future company risk.

Despite theses challenging characteristics, financial reports do contain meaningful information about future company performance. For instance, Cohen et al. (2020) show that large year-over-year changes to the language of company reports indicates a significant negative signal about their future performance and can predict financial variables, such as earnings, profitability, and bankruptcy. While their methods are shown to be effective, they only use simple string similarity

Findings of the Association for Computational Linguistics: EACL 2024, pages 500–512

measures to compare reports.

In a different application, given the detailed information about company business operations contained in these reports, there is an opportunity to identify relationships between companies that can help predict their future market correlation. Public companies are related to each other in various forms and this relationship governs the co-movement of their stock prices. Therefore, the ability to predict that relationship in advance from their reports is valuable to investment managers. These relationships can take various forms, including, having similar products, sharing technologies, or being exposed to the same economic risk factors. (Cohen and Frazzini, 2008; Hoberg and Phillips, 2016; Lee et al., 2019).

In this work, we explore these applications by curating a dataset of paired financial reports and introducing two novels tasks that exploit the cross-document interaction between them to make long-horizon financial predictions. We experiment with a comprehensive set of end-to-end methods to model the interaction between these long financial documents. We find that it is possible to predict stock risk and pairwise correlation solely from text and that methods that allow for a more sensitive and fine-grained interaction between them provide significant benefit. In addition, we find that these text-based models provide considerable value beyond standard financial variables.

We provide a simple yet effective method that can compare arbitrarily long documents at a fine-grained level and identify subtle similarities and differences between them. We train this model end-to-end to allow the model to learn directly from the future financial outcomes associated with each pair of reports, so it can learn to identify subtle, task-specific similarities and differences that are most predictive.

In summary, we make the following contributions:

1. We curate a new dataset of paired company financial reports, containing complex, financial language and cross-document relationships, that we anticipate to be of broad interest to the community (§4, Appendix A).
2. We propose two novel and challenging financial prediction tasks, including forecasting future long-horizon stock risk and pairwise correlation, that both require the ability to recognize subtle similarities and differences between long financial documents (§3). To the best of our knowledge, this is the first work to consider and effectively model the cross-document interactions between paired reports for financial prediction in an end-to-end manner.
3. We systematically investigate and experiment with a comprehensive set of methods for these tasks, including tailored document-level and sentence-level Transformers that achieve strong performance, establishing the state-of-the-art (§5).
4. We demonstrate that while the tasks are challenging and many simple methods perform poorly, it possible to predict company risk and correlation with performance well-above random chance from solely the text of their financial reports by modeling the cross-document relationship at a fine-grained level with tailored pretraining objectives (Table 2).
5. We probe the best performing model through quantitative and qualitative interpretability methods to reveal insight into the underlying task signal (§7).

Broader Impact We hope this work will inspire future research in long document similarity and cross-document modeling by providing a dataset and two challenging tasks, particularly as the context size for LLMs continues to grow. For reproducibility and to advance the study of these research areas, we release the dataset and sample code at: `https://github.com/rosskoval/learn_to_compare_fr/`.

2 Related Work

In the broader NLP literature, there has been great interest recently in extending the context length of Transformer-based language models to be able to efficiently process long documents (Dai et al., 2019; Beltagy et al., 2020; Zaheer et al., 2020; Kitaev et al., 2020; Guo et al., 2022). These methods attempt to approximate full self-attention with more efficient computation and have been shown to excel at long document understanding tasks.

In a related area, long document similarity involves identifying the relationship between two long documents. While semantic similarity has been of interest for a while, most work has focused on short text at the sentence or paragraph-level (Cer et al., 2017). However, semantic similarity at the document-level is more challenging because long documents often contain content spanning multiple

topics and relationships between them may exist at different levels. Despite the difficulty, the problem has varied applications, including citation recommendation, plagiarism detection, coreference resolution, and multi-document summarization. In Zhou et al. (2020), the authors propose a cross-document attention component into HAN (Yang et al., 2016) to enable the comparison between documents at different levels. Further, in Caciularu et al. (2021), the authors consider a similar setting and propose a novel pretraining approach for Cross-Document Language Modeling (CDLM) with a dynamic attention mechanism that allows the model to learn cross-document relationships. They demonstrate that their model has a strong understanding of the relationship between documents and delivers SOTA performance on a variety of multi-document tasks. Other methods have attempted to perform an alignment between related documents at the sentence-level for retrieval applications, but typically pretrain encoders in a self-supervised manner, without finetuning them end-to-end on the target task. (Ginzburg et al., 2021; Di Liello et al., 2022a,b).

2.1 Financial Prediction

In addition to Cohen et al. (2020) which inspired this work, there have been other works that examine using single firm reports for financial forecasting tasks, but primarily in isolation without any comparison to other related documents. For instance, Kogan et al. (2009) extract textual features from the most recent financial report to predict stock volatility, while Koval et al. (2023) directly learn to predict companies' future earnings surprise from the text of their conference call transcripts. Other work in this area has combined textual reports with multimodal data, such as audio, tabular, and financial features to enhance predictions (Sawhney et al., 2020; Feng et al., 2021; Alanis et al., 2022; Mathur et al., 2022).

3 Problem Statement

We propose two novel tasks designed to evaluate the ability to recognize subtle similarities and differences between long financial documents that are predictive of long-horizon financial outcomes. It is important to note that since the reports occur at an annual frequency, we choose target variables at the 1-year horizon, which produces a lot of uncertainty between the forecast and outcome date, and makes these long-horizon prediction tasks particularly challenging. We also believe that the long-horizon requires the ability to capture more intricate, subtle signals than similar short-horizon tasks. In addition, this choice is consistent with prior work (Kogan et al., 2009; Feng et al., 2021; Alanis et al., 2022) and the premise from Cohen et al. (2020) that the text-based signal contained in these reports is related to business risk that potentially can take up to multiple quarters to materialize on company performance.

3.1 Risk Prediction

Risk prediction is a valuable tool for investment managers when constructing a portfolio of financial assets. While there are many measures of financial risk, Maximum Drawdown (MDD) has become an important one, which measures the most significant percentage decline in the value of an asset over a given period of time (Magdon-Ismail and Atiya, 2004; Chekhlov et al., 2004; Gray and Vogel, 2013; Nystrup et al., 2019). Therefore, we choose this as a target variable and use consecutive financial reports as task inputs to learn to identify subtle yet important signals of company risk. Given the Management Discussion and Analysis (MDA) section of the current financial report $D_{i,t}$ for firm i at year t, we wish to learn to compare and contrast it with the previous report $D_{i,t-1}$ to predict whether the company will experience an abnormal decline (MDD) over the next year. While there may signal in only considering the current report, we believe it can be considerably enhanced when contextualized with the previous report to better capture the salient risks factors facing the company. Given daily price data $P_{i,t}$ for company i at time t, we compute the MDD over the next year T as the magnitude of the largest price decline from peak to trough (Drenovak et al., 2022):

$$\text{MDD}_{i,t} = |\min_{t_1 \in \{t,T\}} \frac{P_{i,t_1}}{\max_{t_0 \in \{0,t_1\}} P_{i,t_0}} - 1|$$

For each year, we label companies with the 20% largest drawdowns during each year in the sample as High Risk ($y = 1$) and those in the bottom 80% as Normal Risk (y = 0):

$$y_{i,t} = \begin{cases} 0, & \text{Percentile}(\text{MDD}_{i,t}) < 0.80 \\ 1, & \text{Percentile}(\text{MDD}_{i,t}) \geq 0.80 \end{cases}$$

We carefully choose this task formulation and target variable for a few different reasons. First, it

is common for investment managers and the financial literature to segment portfolios into quintiles, approximating a live trading setting in which investment managers are faced with the decision to exclude certain high-risk stocks from their portfolio at each decision point. Therefore, it has the potential to help them reduce portfolio risk. Second, we carefully selected Maximum Drawdown (MDD) as our target variable because of its ability to capture the effect of extreme events that occur anytime within the time horizon, since it computes the bottom of market prices attained over the horizon, and use the stock's MDD relative to other stocks within the same time period to indicate High Risk stocks to remove the impact of broad market movements and focus on stock-specific events.

Since the dataset is imbalanced by design and High Risk is a clear positive minority class, we use the F1-score as our primary measure of performance evaluation. We believe it accurately reflects the trade-off for an investment manager who is faced with the decision to include or exclude a stock in their portfolio because misclassifying a High Risk stock as Normal Risk is more costly than misclassifying a Normal Risk stock as High Risk.

3.2 Correlation Prediction

In addition to risk, the correlation matrix of stock returns is an equally important measure for the risk management practices of investment managers (Embrechts et al., 2002; Andersen et al., 2007). Therefore, we also propose the task to predict the future correlation between companies' stock prices by learning to identify similarities and differences between their financial reports. We introduce this task to evaluate the ability of the model to capture various forms of relationships between companies.

For computational purposes, we take a subset of the 100 largest companies in our dataset and compute pairwise relationships to generate 4,950 company-company pairs per year. Further, we remove company pairs in which both companies belong to the same industry classification to challenge the model to identify more subtle connections that extend beyond industry keywords, leaving us with 3,836 pairs per year. We measure the relationship as the correlation between their daily stock returns over the next year from their most recent reports. To do so, we compute daily stock returns $r_{i,t}$ from daily prices $P_{i,t}$ for company i at time t and the correlation between their stock returns over the next year from t to T:

$$\text{corr}(r_{i,t}, r_{j,t}) = \sum_{t=1}^{T} \frac{(r_{i,t}-\bar{r_i})(r_{j,t}-\bar{r_j})}{\sqrt{(r_{i,t}-\bar{r_i})^2(r_{j,t}-\bar{r_j})^2}}$$

Then, we normalize them to be $N(0,1)$ within each year to account for the nonstationarity of market correlations over time:

$$y_{i,j,t} = \frac{\text{corr}(r_{i,t},r_{j,t})-\mu_t}{\sigma_t}$$

We use the Spearman Rank Correlation between the model predictions and observed correlations in each year to evaluate model performance. We use these metrics at the year-level rather than aggregated Mean-Squared Error because the relative ranking of the predictions within a given year is more important than their absolute levels given the non-stationarity of market correlations.

4 Data

4.1 Data Acquisition

To curate the dataset, we download preprocessed HTML files of company filings from the Notre Dame Software Repository for Accounting and Finance (Loughran and McDonald, 2011).

We focus our analysis on Section 7A: Management Discussion and Analysis (MDA) section from annual reports of US-based public companies. According to the SEC, this section is intended to provide management's perspective on the business results of the past year and their future prospects for the upcoming year, including information about key business risks. While there are other sections, we choose to focus on the MDA because it reflects a direct communication from company management to shareholders. We use a variety of regular expressions to extract the MDA section and filter the resulting section text for quality in a refined iterative process. We source stock price data from FactSet Prices & Returns API. Please see Appendix A for further details on the data curation process.

4.2 Data Statistics and Task Formulation

To prevent any form of lookahead bias, we temporally partition the dataset according to the report publication date into training (Jan 2010 – Dec 2014), validation (Jan 2015 – Dec 2015), and test (Jan 2016 – Dec 2019) splits. We do not use expanding sample windows for training/validation due to lack of computational resources, but we would expect doing so would improve results

across model types and we validate this hypothesis with the best performing model.

We present an overview of the dataset with summary statistics in Table 1, including document length and linguistic complexity, measured with Gunning FOG Index (Bushee et al., 2018). We confirm that the MDA section of these reports is becoming longer and more complex over time, likely making it increasingly difficult for investors to process the information contained in them.

	Train	Validation	Test
Start Date	Jan-2010	Jan-2015	Jan-2016
End Date	Dec-2014	Dec-2015	Dec-2019
# Samples	8,123	1,617	7,579
# Firms	2,574	1,572	2,170
# Words	13,092	13,455	14,354
# Sents	403	417	426
Linguistic Complexity	10.76	10.98	11.38

Table 1: Summary Statistics of each MDA section in the Financial Report on each sample split.

5 Methods

We explore a comprehensive set of baselines on these novel tasks that range from simple bag-of-words based methods to well-tailored state-of-the-art document-level and sentence-level Transformer-based models, including both generic and domain-adapted versions of each.

5.1 Simple Baselines

First, we establish a variety of simple baselines that indicate the difficulty of the task. **BOW + Sim + Linear** is solely based on the similarity between the reports using TF-IDF weighted, bag-of-words features while **BoW + Linear** concatenates their features together and passes them to a linear classifier. We also include a pretrained financial sentiment classifier **FinBERT-Sent + Linear** (Araci, 2019) applied at the sentence-level (Alanis et al., 2022):

$$\text{FinBERT-Sent} = \frac{\text{\#PositiveSentences - \#NegativeSentences}}{\text{\#TotalSentences}}$$

The results of this baseline clearly distinguish the Risk task from traditional sentiment analysis.

Additionally, given that the positive autocorrelation of risk is well documented in the financial literature (Kambouroudis et al., 2016), we provide a simple autoregressive time-series baseline **AR(1) + Linear** that fits a linear classifier on the 1-year trailing value. While the resulting performance is below that of the best text-based models, it is important to note that the signal contained in the text is largely distinct from and complementary to it (corr < 0.20). Finally, we also include a purely company financial-based linear classifier **FinVar + Linear** with 10 common accounting and stock-price based financial variables (e.g. valuation, profitability, volatility, price momentum, etc.) to serve as a traditional financial baseline (Alanis et al., 2022). Please see A for more details on the variables used.

5.2 Document-Level Transformers

We consider two approaches to predicting the relationship between two long documents at the document-level, including the Bi-Encoder (BE) and the Cross-Encoder (CE).

5.2.1 Document Encoder

First, we select our primary document encoder to be the Longformer-base because it has been shown to excel at document matching (Caciularu et al., 2021). The model applies a combination of local and global attention to efficiently approximate the full attention matrix.

For the Risk Prediction task, we provide single document baselines that only make use of the current report D_t (**Longformer-Curr**) and previous report D_{t-1} (**Longformer-Prev**), respectively, as well as one that performs a soft "diff" operation between them, only extracting those not contained in the previous report (**Longformer-Diff**), to further justify the use of more sophisticated cross-document methods. We find that several variations of the "diff"-based approach perform worse than just using the current report, which we conjecture is for two reasons. First, the changes are subtle and difficult to identify using manual heuristics. Second, the salient sentences require the surrounding context to effectively contextualize the meaning.

5.2.2 Cross-Encoder (CE)

We also experiment with the Cross-Encoder approach (**Longformer-CDLM-CE**) of concatenating the document text together and use the CDLM-pretrained Longformer model from Caciularu et al. (2021). This approach implicitly interacts the tokens between the documents via the local/global attention mechanism, but the granularity of the interaction may be limited because attention is limited to a local window and special global tokens.

We follow the CDLM-framework and allocate global attention to the first [CLS] token and special document separator tokens <doc-s> and </doc-s>. We extend the maximum length of the model to 8192 tokens by copying over the position embeddings, and then concatenating the first 4096 tokens of each document together.

$$\text{CE}(D_i, D_j) = g([D_i; D_j])$$

5.2.3 Bi-Encoder (BE), Document-Level

Second, we experiment with encoding each document independently and then passing them through a 1-hidden layer MLP for interaction via concatenation of the document embeddings, known as a Bi-Encoder approach (**Longformer-BE**). Consider the document encoder g and related documents D_i and D_j that are encoded as $g(D_i) = E_i$ and $g(D_j) = E_j$, respectively:

$$\text{BE}(E_i, E_j) = \text{MLP}([E_i; E_j; |E_i - E_j|)$$

This interaction function was inspired by Reimers and Gurevych (2019) for sentence-level semantic similarity and we continue to include the absolute value difference term to impose the inductive bias that encourages the model to compare and contrast documents.

5.3 Sentence-Level Transformers

We also experiment with methods that operate on the sentence-level. Since the related documents have a different number of sentences in varying order, we explore a simple yet effective method to perform a soft-alignment between them.

5.3.1 Sentence Encoder

First, we divide each document into sentences and encode each sentence $s_i \in S_i$ and $s_j \in S_j$, using a pretrained sentence encoder f to get sentence embeddings $e_i \in E_j$ and $e_j \in E_j$, in each report, respectively. This model produces contextualized embeddings of all tokens and we extract the last hidden state of the first [CLS] token as the sentence representation (Devlin et al., 2019).

Since the task requires the detection of subtle similarities and differences between topically similar text, it is important to have a sentence encoder that is well-attuned to semantic similarity and the financial domain. Therefore, we explore both pretrained encoders, such as **SBERT** (Reimers and Gurevych, 2019) and **FinBERT** (Huang et al., 2022), as well as the **DiffCSE** (Chuang et al., 2022) framework to pretrain a sentence encoder on our in-domain corpus. DiffCSE improves upon the SimCSE (Gao et al., 2021) framework, which uses stochastic dropout-based augmentations as positive pairs and in-batch negatives with contrastive learning, by incorporating an additional Replaced Token Detection (RTD) loss that conditions upon the original sentence representation to predict the location of randomly replaced tokens that were generated by a fixed masked language model. This additional objective has been shown to make the encoder more sensitive to small yet important differences in sentences.

5.3.2 Cross-Document Sentence Alignment (CDSA)

The IR literature suggests that methods with token-level interactions provide a more fine-grained and powerful approach for query-document similarity tasks than those that operate at the document-level (Khattab and Zaharia, 2020; Zhou et al., 2020). With this in mind, we explore a simple yet effective extension of this approach to align and compare long financial reports at the sentence-level, which we denote as Cross-Document Sentence Alignment (**CDSA**).

To do so, we employ a cross-attention mechanism between the sentence embeddings of both documents to perform a soft-alignment, inspired by encoder-decoder attention (Bahdanau et al., 2014; Vaswani et al., 2017), which operates at a token-level. This mechanism creates a unique and corresponding context vector for each sentence in the focal report by attention weighting all sentences in the related report, and represents the portion of information of that sentence that is contained in the other report. We apply this in both directions, for each sentence embedding $e_i \in E_i$ across sentences embeddings E_j, and for each sentence $e_j \in E_j$ across sentence embeddings E_i:

$$c_i = \sum_{e_j \in E_j} \alpha_{i,j} e_j$$

$$c_j = \sum_{e_i \in E_i} \alpha_{j,i} e_i$$

where the attention weight α is given by softmax, dot-product attention (Vaswani et al., 2017).

To adapt the document-level Bi-Encoder approach BE to the sentence-level, we can compare each sentence embedding e_i, e_j with the corresponding soft-aligned context vector c_i, c_j from

CDSA using a similar interaction function:

$$\text{BES}(e_i, c_i) = \text{MLP}([e_i; c_i; |e_i - c_i|])$$

Then, we conduct simple mean pooling over all sentence-level MLP outputs:

$$m(E_i) = \frac{1}{|E_i|} \sum_{e_i \in E_i} \text{BES}(e_i, c_i);$$

$$m(E_j) = \frac{1}{|E_j|} \sum_{e_j \in E_j} \text{BES}(e_j, c_j)$$

Finally, we concatenate the pooled outputs from both reports $m(E_i)$ and $m(E_j)$ and pass them through a classifier for prediction:

$$\hat{y} = \sigma([m(E_i); m(E_j)])$$

This mechanism allows for the detection of similarities and differences across each sentence in both reports.

5.4 Domain Adaptive Pretraining (DAPT)

Domain adaptation is important to the success of using pretrained language models for out-of-distribution text (Han and Eisenstein, 2019; Gururangan et al., 2020). Since we believe our tasks require a nuanced understanding of financial language, we conduct domain-adaptive pretraining (**DAPT**) for all of the baseline models. To do so, we aggregate a collection of 30K paired annual reports published between 2000 and 2009, prior to the start of the training data to prevent any form of data leakage, and create an in-domain pretraining corpus for all forms of DAPT in this work for fair comparison across model types.

5.4.1 Document-Level

For the document-level models with a Longformer backbone, we conduct DAPT across the following different pretraining objectives: long context MLM (Beltagy et al., 2020) denoted as **Longformer-BE + DAPT w/ MLM**, CDLM (Caciularu et al., 2021) with pairs of consecutive reports (**Longformer-CE + DAPT w/ CDLM**); and follow the same pretraining settings and hyperparameters as Beltagy et al. (2020) and Caciularu et al. (2021), respectively.

We also adapt the DiffCSE pretraining framework designed for short-context models, to the Longformer backbone model (**Longformer-CE + DAPT w/ DiffCSE**) for more sensitive document representations by prepending and assigning global attention to the original document embedding in the RTD objective to encourage the model to use that information to predict the replaced tokens.

5.4.2 Sentence-Level

We also use this corpus for pretraining a more domain-adapted and sensitive sentence encoder from the RoBERTa checkpoint using the DiffCSE framework (**CDSA-FinDiffCSE**) but limit the size to 10M sentences for computational purposes, and use the same pretraining settings and hyperparameters in Chuang et al. (2022). We expect this pretraining step to be able to better differentiate topically similar yet semantically different financial language.

Finally, since the validation data (2015) and last year of the test data (2019) are 4 years apart, we experiment with an expanding window training/validation approach (**CDSA-FinDiffCSE + Expanding**) to allow the model to access more recent data and simulate a production trading environment. However, we only do this for the best performing model because it is not computationally feasible to do for all models. We also include a simple multimodal approach (**CDSA-FinDiffCSE + AR(1)**) that fits a linear combination between the predictions of the CDSA-FinDiffCSE and AR(1) models. Please see Appendix A for further details.

5.4.3 Implementation Details

Finally, we train all of these baseline models on each financial prediction task with binary cross-entropy loss and mean-squared error for the Risk and Correlation prediction tasks, respectively. For fair comparison across model types, we only consider the first 4096 tokens in each report; see Appendix A for further implementation details.

6 Experimental Results and Analysis

The results in Table 2 highlight the challenging nature of both tasks, but **we find broad consistency in the relative performance results across them**, with CDSA-FinDiffCSE performing the best in both with statistical significance, and improving considerably from expanding training data. This result provides evidence that while the tasks are distinct, they both require the ability to recognize subtle similarities and differences between long documents at a fine-grained level, and this ability is directly correlated with the relative ranking of model performance.

In general, **we find that the sentence-level methods generally perform better than the document-level methods**, which we conjecture is because by they allow for a more fine-grained

Model	# Params	Risk Prediction					Correlation Prediction				
		$F1_{2016}$	$F1_{2017}$	$F1_{2018}$	$F1_{2019}$	Avg	ρ_{2016}	ρ_{2017}	ρ_{2018}	ρ_{2019}	Avg
Minority Class All-1	0	0.33	0.33	0.33	0.33	0.33	-	-	-	-	-
BoW + Sim + Linear	2	0.36	0.35	0.35	0.34	0.35	0.10	0.16	0.07	0.06	0.10
BoW + Linear	100K	0.41	0.39	0.38	0.37	0.38	0.14	0.20	0.19	0.19	0.18
FinBERT-Sent + Linear	2	0.38	0.38	0.38	0.38	0.38	-	-	-	-	-
AR(1) + Linear	2	0.42	0.40	0.44	0.40	0.42	0.25	0.25	0.25	0.26	0.25
FinVar + Linear	11	0.45	0.43	0.48	0.45	0.45	-	-	-	-	-
Longformer-Prev	152M	0.42	0.43	0.40	0.35	0.40	-	-	-	-	-
Longformer-Curr	152M	0.48	0.47	0.47	0.45	0.47	-	-	-	-	-
Longformer-Diff	152M	0.44	0.43	0.43	0.39	0.43	-	-	-	-	-
Longformer-BE	152M	0.48	0.47	0.47	0.44	0.47	0.11	0.24	0.19	0.08	0.15
Longformer-CDLM-CE	152M	0.51	0.48	0.50	0.44	0.48	0.12	0.26	0.20	0.13	0.18
Longformer-BE + DAPT w/ MLM	152M	0.49	0.45	0.49	0.45	0.47	0.16	0.25	0.28	0.24	0.23
Longformer-BE + DAPT w/ DiffCSE	152M	0.52	0.48	0.49	0.46	0.49	0.26	0.34	0.29	0.24	0.28**
Longformer-CE + DAPT w/ CDLM	152M	0.53	0.49	0.50	0.46	0.50	0.22	0.30	0.31	0.24	0.27
CDSA-RoBERTa	128M	0.51	0.48	0.48	0.42	0.47	0.27	0.24	0.24	0.19	0.24
CDSA-SBERT	115M	0.54	0.52	0.51	0.44	0.50	0.28	0.31	0.27	0.18	0.26
CDSA-DiffCSE	128M	0.51	0.52	0.52	0.47	0.51	0.28	0.33	0.28	0.19	0.27
CDSA-FinBERT	128M	0.53	0.51	0.53	0.48	0.51*	0.30	0.32	0.25	0.17	0.26
CDSA-FinDiffCSE	128M	0.55	0.54	0.52	0.51	0.53*	0.30	0.33	0.32	0.27	0.31**
CDSA-FinDiffCSE + AR(1)	128M	0.58	0.56	0.57	0.53	0.56	0.37	0.45	0.46	0.33	0.40
CDSA-FinDiffCSE + Expanding	128M	0.55	0.55	0.59	0.57	0.56	0.30	0.40	0.36	0.31	0.34

Table 2: Main Results - Model performance on the test set of the Risk and Correlation Prediction task. All performance numbers are reported in decimal and the top 2 models within each task are bolded. "-BE" indicates Bi-Encoder while "-CE" indicates Cross-Encoder document-level models as defined in §5. "+ Expanding" indicates that expanding training/validation sample windows was used. "+ AR(1)" indicates that a linear combination of the predictions was fit between the CDSA-FinDiffCSE and AR(1) model. *, ** indicates the performance of the best model is statistically better ($p < 0.01$) than that of the second best model according the Wilcoxon Signed-Rank Test.

interaction between the document sentences before any document-level pooling. We find this effect to be more pronounced on the Correlation Prediction task, especially when the Longformer base model is not pretrained for semantic similarity. This suggests that despite the extensive, language modeling-based pretraining process of the Longformer model, it does not produce strong document embeddings without finetuning.

However, we find that our long context adaptation of the DiffCSE pretraining framework for the Longformer is well-suited for generating fine-grained document embeddings, suggesting that this is a promising direction for future work.

Relatedly, **we find that pretrained models not adapted to the financial domain or pretrained with semantic similarity objectives struggle to learn the subtle task signals**. However, we observe a significant improvement across most models after DAPT, suggesting that the task requires a nuanced understanding of financial language.

For both tasks, we find that a simple multi-modal model **CDSA-FinDiffCSE + AR(1)** improves performance, particularly for the Correlation Prediction task which exhibits stronger autocorrelation. We conjecture their complementary nature is partly due to the fact that historical market patterns captures the persistence of past behavior while the text-based models identify the catalyst that causes novel behavior, suggesting the text-based methods could serve as a valuable tool to augment traditional risk management practices. However, we leave it to future work to explore more sophisticated methods to incorporate tabular data into text-based models.

7 Model Interpretability and Analysis

7.1 LM Sensitivity Analysis

To further understand model behavior on the Risk Prediction task, we perform a simple interpretability test using the LM financial dictionary (Loughran and McDonald, 2011) and the predictions of the best performing model (CDSA-FinDiffCSE). We provide an overview of the summary statistics of the dictionary and results in Table 3. To do so, we extract the model predicted probabilities, and regress them onto the changes in the proportion of LM dictionary words between the current and previous report to understand their linear relationship.

In Table 3, we observe that the model's predic-

Category	# words	% words	% sentences	coeff	p-value
Δ Positive	347	0.55	13.59	-4.06	0.04
Δ Negative	2345	1.32	24.92	4.12	0.04
Δ Uncertain	297	1.36	30.34	4.56	0.13
Δ Litigious	903	0.59	13.10	1.05	0.18
Δ Constraining	184	0.57	14.23	6.12	0.05
Δ Strong Modal	19	0.23	6.53	11.46	0.11
Δ Weak Modal	27	0.59	15.20	0.58	0.91

Table 3: Linear Regression of the model predictions onto the YoY changes in LM financial sentiment variables.

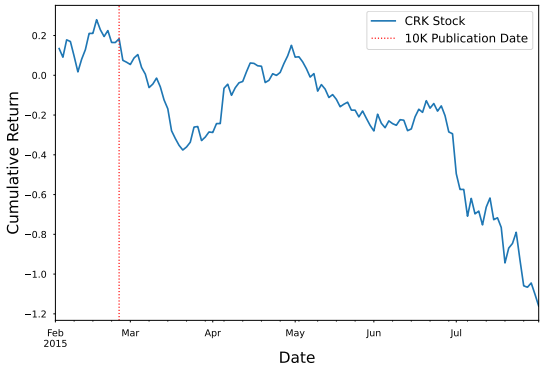

Figure 2: Stock Price of CRK following the publication of the 2015 Annual Report that was identified by the best performing model as High Risk based off changes from the 2014 Annual Report.

tions for High Risk are negatively associated with increases in positive financial sentiment, and positively associated with increases in negative, constraining, and litigious financial sentiment. While some variables are statistically significant and the results are economically intuitive, the linear model has an adjusted R^2 of just 3.4%, indicating that the trained model is capturing more powerful features than only simple changes in LM sentiment. We also note the positive correlation with increases in strong modal words is consistent with Loughran and Mcdonald (2011), who find that firms with higher proportions of strong modal words in their quarterly reports are more likely to subsequently report material weakness in their accounting controls, which is likely a strong signal for increases in the likelihood of future High Risk behavior.

7.2 Case Study and Qualitative Analysis

We conduct a case study of the reports of Comstock Resources Inc, referenced (CRK) in Figure 1. We find that the report scores highly as High Risk by the best performing CDSA model and correctly identifies the salient risky sentences, as measured via the largest L_2 norms in the $|s_i - c_i|$ term, which we highlighted in the exhibit. As shown in Figure 2, the company stock price experienced a precipitous drawdown of more than 100% in the 6 months following the release of this report. We find that the model was able to detect subtle yet important changes in the text that predicted a large drawdown months before it occurred.

8 Conclusion

We curate a large-scale corpus of paired annual financial reports and introduce two novel benchmarks that require modeling complex, cross-document interactions between long documents. We methodically investigate a comprehensive set of methods that are well-attuned to the task, establishing the state of the art. Through analysis of the experimental results and use of interpretability methods, we reveal insights into the underlying task signals. We hope our contributions inspire further research in this important area.

Limitations

Our experiments demonstrate that it is possible to analyze and compare the financial reports of public companies to predict future company risk and correlation with performance that is well above random chance. However, we acknowledge that the Risk Prediction task is formulated as a classification setting so the results do not necessarily directly translate to a live trading setting and that the absolute values of the performance numbers in the Correlation Prediction task are relatively low so we leave it to future work to assess their utility in real-world portfolio management.

Ethics Statement

We acknowledge that our 10K Annual Financial Report dataset contains English reports from the largest US-based companies so it is possible that some populations may be underrepresented in this sample. We plan to extend this work to international companies and financial reports written in other languages in the future.

Acknowledgements

We would like to thank AJO Vista and FactSet for providing access to and permission to release the data. The authors are solely responsible for the

content and views expressed in this publication and do not reflect those of the affiliated institutions.

References

Emmanuel Alanis, Sudheer Chava, and Agam Shah. 2022. Benchmarking machine learning models to predict corporate bankruptcy. *arXiv preprint arXiv:2212.12051*.

Torben G Andersen, Tim Bollerslev, Peter Christoffersen, and Francis X Diebold. 2007. Practical volatility and correlation modeling for financial market risk management. In *The risks of financial institutions*, pages 513–548. University of Chicago Press.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Brian J Bushee, Ian D Gow, and Daniel J Taylor. 2018. Linguistic complexity in firm disclosures: Obfuscation or information? *Journal of Accounting Research*, 56(1):85–121.

Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew E Peters, Arie Cattan, and Ido Dagan. 2021. Cdlm: Cross-document language modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2648–2662.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

Alexei Chekhlov, Stanislav Uryasev, and Michael Zabarankin. 2004. Portfolio optimization with drawdown constraints. In *Supply chain and finance*, pages 209–228. World Scientific.

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljačić, Shang-Wen Li, Scott Yih, Yoon Kim, and James Glass. 2022. Diffcse: Difference-based contrastive learning for sentence embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4207–4218.

Lauren Cohen and Andrea Frazzini. 2008. Economic links and predictable returns. *The Journal of Finance*, 63(4):1977–2011.

Lauren Cohen, Christopher Malloy, and Quoc Nguyen. 2020. Lazy prices. *The Journal of Finance*, 75(3):1371–1415.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988.

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Luca Di Liello, Siddhant Garg, Luca Soldaini, and Alessandro Moschitti. 2022a. Paragraph-based transformer pre-training for multi-sentence inference. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2521–2531.

Luca Di Liello, Siddhant Garg, Luca Soldaini, and Alessandro Moschitti. 2022b. Pre-training transformer models with sentence-level objectives for answer sentence selection. *arXiv preprint arXiv:2205.10455*.

Mikica Drenovak, Vladimir Ranković, Branko Urošević, and Ranko Jelic. 2022. Mean-maximum drawdown optimization of buy-and-hold portfolios using a multi-objective evolutionary algorithm. *Finance Research Letters*, 46:102328.

Paul Embrechts, Alexander McNeil, and Daniel Straumann. 2002. Correlation and dependence in risk management: properties and pitfalls. *Risk management: value at risk and beyond*, 1:176–223.

Qi Feng, Han Chen, and Ruohan Jiang. 2021. Analysis of early warning of corporate financial risk via deep learning artificial neural network. *Microprocessors and Microsystems*, 87.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Dvir Ginzburg, Itzik Malkiel, Oren Barkan, Avi Caciularu, and Noam Koenigstein. 2021. Self-supervised document similarity ranking via contextualized language models and hierarchical inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3088–3098.

Wesley R Gray and Jack Vogel. 2013. Using maximum drawdowns to capture tail risk. *Available at SSRN 2226689*.

Mandy Guo, Joshua Ainslie, David C Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. Longt5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 724–736.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4238–4248.

Gerard Hoberg and Gordon Phillips. 2016. Text-based network industries and endogenous product differentiation. *Journal of Political Economy*, 124(5):1423–1465.

Allen H Huang, Hui Wang, and Yi Yang. 2022. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*.

Dimos S Kambouroudis, David G McMillan, and Katerina Tsakou. 2016. Forecasting stock return volatility: A comparison of garch, implied volatility, and realized volatility models. *Journal of Futures Markets*, 36(12):1127–1163.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*.

Shimon Kogan, Dimitry Levin, Bryan R. Routledge, Jacob S. Sagi, and Noah A. Smith. 2009. Predicting risk from financial reports with regression.

Ross Koval, Nicholas Andrews, and Xifeng Yan. 2023. Forecasting earnings surprises from conference call transcripts. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8197–8209, Toronto, Canada. Association for Computational Linguistics.

Charles MC Lee, Stephen Teng Sun, Rongfei Wang, and Ran Zhang. 2019. Technological links and predictable returns. *Journal of Financial Economics*, 132(3):76–96.

Tim Loughran and Bill McDonald. 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of finance*, 66(1):35–65.

Tim Loughran and Bill Mcdonald. 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *Journal of Finance*, 66.

Malik Magdon-Ismail and Amir F Atiya. 2004. Maximum drawdown. *Risk Magazine*, 17(10):99–102.

Puneet Mathur, Mihir Goyal, Ramit Sawhney, Ritik Mathur, Jochen L Leidner, Franck Dernoncourt, and Dinesh Manocha. 2022. Docfin: Multimodal financial prediction and bias mitigation using semi-structured documents. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1933–1940.

Peter Nystrup, Stephen Boyd, Erik Lindström, and Henrik Madsen. 2019. Multi-period portfolio selection with drawdown control. *Annals of Operations Research*, 282(1-2):245–271.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Ramit Sawhney, Piyush Khanna, Arshiya Aggarwal, Taru Jain, Puneet Mathur, and Rajiv Shah. 2020. Voltage: Volatility forecasting via text audio fusion with graph convolution networks for earnings calls. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 8001–8013.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Ho Chung Wu, Robert Wing Pong Luk, Kam Fai Wong, and Kui Lam Kwok. 2008. Interpreting tf-idf term weights as making relevance decisions. *ACM Transactions on Information Systems*, 26.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. volume 2020-December.

Xuhui Zhou, Nikolaos Pappas, and Noah A. Smith. 2020. Multilevel text alignment with cross-document attention.

A Appendix

A.1 Data Curation

To extract the MDA section from the HTML files, we begin by searching for strings that begin with "Item 7: Management Discussion and Analysis" and conclude with "Item 7A: Quantitative and Qualitative Disclosures", as well as other variations of these patterns in a refined and iterative process to achieve the best coverage. This process required an extensive amount of text processing that was required to extract the relevant sections required many different regular expressions, extensive trial-and-error, and a significant amount of manual quality filtering. We next pair reports for companies based on their fiscal calendar and reporting dates, allowing for delays and differences in publication dates. Finally, we filter the section text for validity and quality, such as ensuring each text has at least 500 words. We also choose focus on annual rather than quarterly reports because their formatting is more standardized and consistent.

A.2 Text-Based Baseline Models

We use Scikit-learn develop the BoW models. We apply the following text preprocessing steps to create input features: remove stop words and rare words; create both unigrams and bigrams; and apply Term Frequency-Inverse Document Frequency weighting (TF-IDF; Salton and Buckley, 1988; Wu et al., 2008).

We develop the neural models in PyTorch and source pretrained checkpoints from HuggingFace. We perform several variations of the Longformer-Diff model over different ways to measure sentence-sentence similarity, only reporting the configuration with the best result on the validation set in Main Results for brevity, including Jaccard Similarity and Cosine Similarity between SBERT pretrained sentence embeddings. We also vary the cutoff threshold over $\{0.10, 0.25, 0.50, 0.75, 0.90\}$ to define a sentence in the current report that is sufficiently different from those in the previous report.

We use an Expanding training/validation window for the best performing model (CDSA-FinDiffCSE + Expanding) to simulate a live trading setting in which we do walk-forward prediction by expanding the training and validation set by 1-year as we predict on the next year of the test set. For instance, when we make predictions on the test set for 2018, we use training data from 2010-2016 and 2017 as validation data. We only due this for the best performing model to provide a proof-of-concept because it is too computationally expensive to do for all models.

A.2.1 Financial-Based Baseline Model

We select 10 commonly used market price and accounting-based financial variables available at the time of the report from the literature (Alanis et al., 2022), including dividend yield, valuation, growth, profitability, medium-term price momentum, short-term price reversal, volatility, leverage, liquidity, and size. This baseline is not intended to be comprehensive in including all possible relevant financial variables to the prediction task but rather to serve as a reasonable baseline approximating common risk factor models employed in the financial industry against which to reference and compare the value of text-based models. There may be other relevant financial variables such as those source from the options or corporate credit market to which we do not have access and is out of the scope of this text-based focused work.

A.3 Training Details and Hyperparamter Tuning

All neural network-based experiments are performned on a single Tesla A100 GPU with 40GB in memory and use AdamW to optimize all parameters. We tune the hypeparamters with a grid search over learning rates $\in \{3e-5, 5e-5, 7e-5\}$, weight decay $\in \{1e-3, 1e-2\}$ and batch size $\in \{32, 64\}$, based off validation set performance. We train all models for 10 epochs and select the best checkpoint based off validation set performance for test evaluation. For computational constraints, we use mixed precision training and gradient checkpointing to satisfy GPU memory constraints. It takes approximately 30 minutes per epoch of supervised finetuning for the sentence-level models and 60 minutes per epoch for the document-encoder models.

A.4 DAPT Pretraining Details

We conduct the DAPT process for the document-level, Longformer backbone models for a maximum of 25K training steps or until the loss on the validation set increases, using the same hyperparameter configuration and settings as Caciularu et al. (2021). This pretraining process takes multiple days of run time for each framework and indicates the difficulty of pretraining these Efficient Transformers models on domain relevant text.

We conduct the DAPT process for the sentence encoder with the DiffCSE framework for a maximum of 100K training steps or until validation loss increases. For the Longformer DAPT w/ Diffcse model, we use Longformer base as the fixed generator (masked language model) model because there are no widely accepted distilled or smaller versions. For both sentence and document encoders, we tune the RTD loss weight in the DiffCSE objective over $\{0.01, 0.05, 0.10, 0.50\}$ according to validation set performance. Please see Chuang et al. (2022) for more details on the framework.