Exploring the relationship between descriptions of research and long term company performance

November 20, 2023

1 Related Work

1.1 NLP for financial texts

While structured numerical data is the predominant form of financial data, extracting contextual financial information from text and documents, using natural language processing (NLP) methods, has become a rapidly developing research field. Some of the commonly used financial texts include company news reported by newspaper and journals, [7,8] earnings call transcript [1,5,11,14], expert analyst reports [13,15] and SEC filings. [18] The financial analysis is usually framed as a text classification problem, and a mixture of traditional and deep learning based methods have been used. Examples include predicting the Lipper Global category of ETFs [18], short and long term stock movement [5, 11] and categorization of digital strategy. [1, 14]

While non-deep learning based NLP methods are domain agnostic, deep learning based large language models often perform better when they are trained or fine-tuned on a specific corpus. FINBERT [2] and BloombergGPT [19] are two large language models trained specifically on a finance corpus and have shown improved performance on a wide range of downstream finance problems.

1.2 Relationship between R&D and company performance

The relationship between a company's investment in research and development and how it performs has been widely investigated in the finance and economics literature. R&D spending has been closely associated with growth in sales and revenue (but not profitability) for top US companies across various industries in a ten-year period during the 1970s and 80s. [12] A targeted sector-based analysis showed a

positive association between R&D growth and sales growth for high R&D intensity sectors such as software and pharmaceuticals, and the high profitability of highly R&D-intensive sectors shows that successful R&D is able to directly boost business performance. [16] Investments in R&D have also been associated with delayed returns in the stock market. [6] Outside the US, similar analysis has shown the existence of a positive association of R&D spending with firm revenue and profit in Vietnam [17] and stock market performance in Turkey. [4]

2 Data Extraction and filtration

The financial text used in this work is the Form 10-K report filed by public companies to US Security and Exchange Commission (SEC) every year. The K-10 form provide a comprehensive description of a company's economic activity during the past financial year and also includes projection into the future in the form of analysis of the company's position in the market and the relevant risks associated with its position. K-10 forms can be retrieved through SEC's Electronic Data Gathering, Analysis, and Retrieval system (EDGAR). We use EDGAR-CORPUS [10], which is a corpus of all K-10 filings from 1993 to 2020 as our primary source of textual data. In total, the corpus contains 6.5 billion tokens from 38,009 companies.

We have incorporated all companies that have K-10 filing for our base year and our target year between 2010 and 2022. A company that has 10K filings only for year 2012, 2013 and 2015 will be included in our data set when training model with base year 2012 and target years 2013 and 2015, it will also be used to train a model with base year 2013 and target year 2015 while being exuded from other models with different base and target year for which there were no filings. Each model is trained on 6800 to 8200 companies based on availability of 10-K filings for the year. To extract information relevant only to research and development, we curate a list of 39 keywords that are commonly applicable in a R&D context, and filter the K-10 text to only keep sentences that have at least one of these keywords. This amounted to a total of 36,564 filings (one filing represents a particular company's K-10 form for a particular year), with an average of 252 sentences (9232 tokens) for each filing. This is our *base* corpus.

We also create a *future looking* corpus, containing only future looking statements. The future looking corpus is further filtered from the base corpus by only including sentences that have a "future looking statement" and project into the future. To do this, we utilize word lists curated by previous studies and use a combination of them. [3, 9]. The future looking corpus is about half the size of the base corpus: 36,564 filings with an average of 120 sentences (5278 tokens) per filing.

3 Setting up the research question

We treat the problem of predicting company performance from K-10 filings as a text classification problem. Given a set of text from K-10 filings, we will experiment with various text vectorization methods and machine learning models that generate a label output. Some ways we labelize the output to encode company performance are given below.

3.1 Revenue CAGR from 2009-2021

We extract the revenue per year of the company from SEC's archive of all company facts[1]. The archive contains financial information of companies across multiple years compiled from the Extensible Business Markup Language (XBRL) files submitted by the company to SEC. To extract revenue from these files, we use three concepts from us-gaap taxonomy that relate to revenue. We only extract annual data included in a K-10 form for the sake of data consistency.

We calculate the revenue compound annual growth rate as $\text{CAGR} = \left(\frac{R_f}{R_s}\right)^{\frac{1}{f-s}}$, where R_f and R_s are the company's revenue in the target year f and base year s, with $f > s$.

We have created the three labels that help distinguish the extraordinarily (good or bad) companies from the norm. We employed quartiles derived from the CAGR of each respective year to establish the following labels:

- *Label 0* includes companies where CAGR $<$ Q1;
- *Label 1*, which includes companies where Q1 $<$ CAGR $<$ Q3;
- *Label 2*, which includes companies where CAGR $>$ Q3.

3.2 Change in ranking of R&D spending growth

Additionally, we also aim to relate amount of research spending with how the company reports its research in text. We obtain the raw numerical R&D spending of a company in the same way as we collected revenue. R&D is reported by a significantly less number of companies, so we dropped some companies for this task. In total, only about 914 companies were analyzed.

[1] http://www.sec.gov/Archives/edgar/daily-index/xbrl/companyfacts.zip

The main goal is to predict the change in growth of research spending of the company over the 2010s. In order to do so, first, we calculate the rankings of company based on their CAGR of their research spending from 2009-2015 (r_1). Then, we will separately calculate the ranking of companies based on CAGR of research spending from 2015-2020 (r_2). If $r_2 > r_1$, the company gets a positive label. In other words, if a company's R&D spending Due to the nature of the task construction, the labels are more balanced than the previous task.

4 Numerical correlations

For a sanity check and to ensure labels for our research questions are valid, we construct a simple correlation between research spending and various company performance metrics. We found a consistent correlation between 0.43 and 0.67 when revenue and research spending for each year were correlated. Additionally, to account for the future (potential) revenue arising from current R&D spending, we correlate the research spending for a year against revenue staggered by anything from 2–10 years.With this staggered data, we found a consistent correlation from 0.4-0.65.

Additionally, we also conduct a Mann-Whitney test to check that the statistical distributions of the R&D spending of three labels generated by CAGR of revenues are valid. All three test results show $p < 0.001$. Additionally, we use Cliff's δ as a measure of effect size. We find an δ of 0.5 between labels 0 and 1, an δ of 0.2 between label 0 and label 2 and an δ of -0.4 between label 1 and label 2. Note that this shows that while the distributions of R&D spendings between the three labels are statistically significant from each other, the probability of the R&D spending of label 1 is higher than that of label 2, which is somewhat counterintuitive to our idea of a linear relationship between R&D and revenue growth since label 2 companies have higher CAGR of revenue than label 1 companies.

5 TF-IDF based text classifier

TF-IDF score is a simple statistic that encodes the relevance of a particular word to a document in a corpus. TF-IDF scores are given by calculating term frequency (TF=$\frac{n_{t,d}}{\sum_{t' \in d} n_d}$) and the inverse document frequency (IDF= $\log \frac{|D|}{|d \in D : t \in D|}$), where t is a term d is a document and D is the entire corpus, or collection of documents. For a particular term t, in a document d, the TF-IDF score is the simple product of TF and IDF.

In our corpus, each filing acts as a document, and is vectorized using TF-IDF scores. Using the vectors, we use Linear Regression to classify the documents. We use a standard 10-fold cross-validation and train-test split with the test size of 10%. To prevent data leakage, we perform the test-train set on the companies first, ensuring that no filing from a company in the test set is in the train set and vise versa. The results for classification of revenue CAGR are given below:

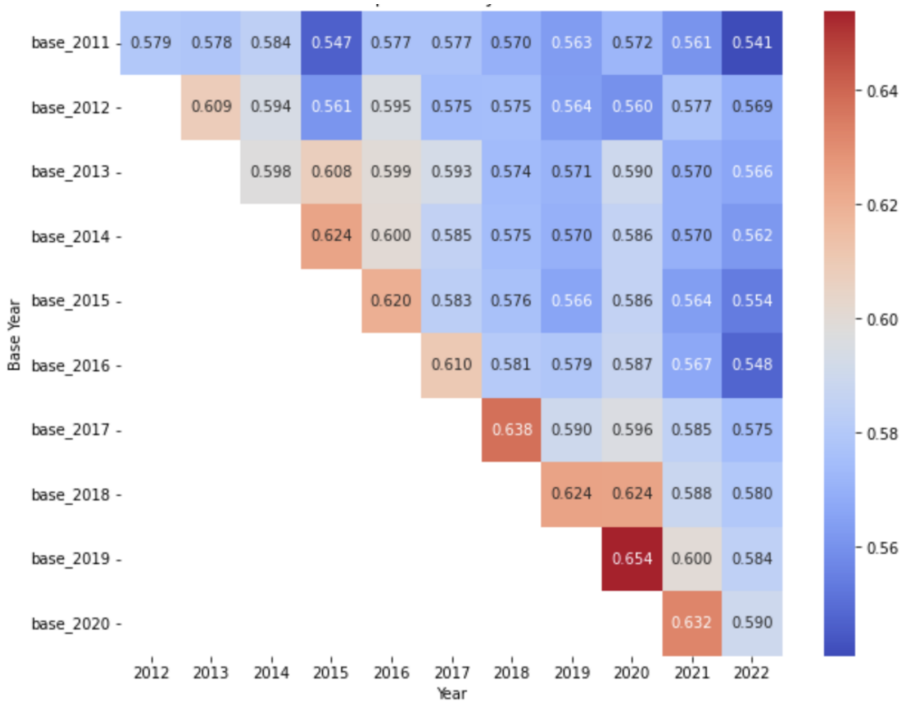

Figure 1: Heat of Accuracy TFIDF Base Corpus

Target	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
base_2011	0.579	0.578	0.584	0.547	0.577	0.577	0.570	0.563	0.572	0.561	0.541
base_2012	-	0.609	0.594	0.561	0.595	0.575	0.575	0.564	0.560	0.577	0.569
base_2013	-	-	0.598	0.608	0.599	0.593	0.574	0.571	0.590	0.570	0.566
base_2014	-	-	-	0.624	0.600	0.585	0.575	0.570	0.586	0.570	0.562
base_2015	-	-	-	-	0.620	0.583	0.576	0.566	0.586	0.564	0.554
base_2016	-	-	-	-	-	0.610	0.581	0.579	0.587	0.567	0.548
base_2017	-	-	-	-	-	-	0.638	0.590	0.596	0.585	0.575
base_2018	-	-	-	-	-	-	-	0.624	0.624	0.588	0.580
base_2019	-	-	-	-	-	-	-	-	0.654	0.600	0.584
base_2020	-	-	-	-	-	-	-	-	-	0.632	0.590

Table 1: TFIDF accuracy for different base and target years using base corpus

Results for future corpus are given below:

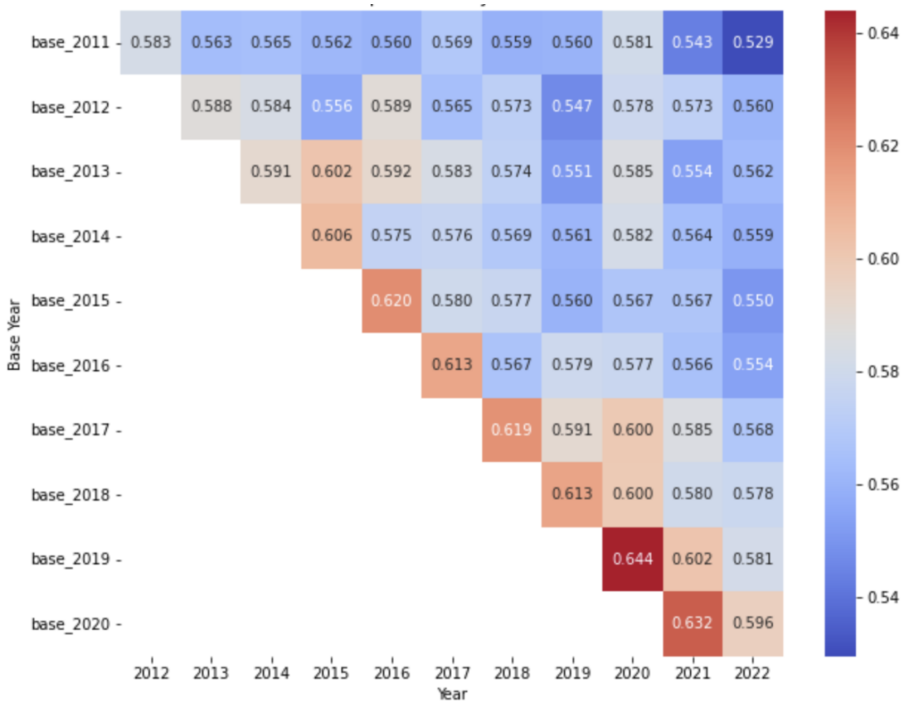

Figure 2: Heat of Accuracy TFIDF Future Looking Corpus

Target	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
base_2011	0.583	0.563	0.565	0.562	0.560	0.569	0.559	0.560	0.581	0.543	0.529
base_2012	-	0.588	0.584	0.556	0.589	0.565	0.573	0.547	0.578	0.573	0.560
base_2013	-	-	0.591	0.602	0.592	0.583	0.574	0.551	0.585	0.554	0.562
base_2014	-	-	-	0.606	0.575	0.576	0.569	0.561	0.582	0.564	0.559
base_2015	-	-	-	-	0.620	0.580	0.577	0.560	0.567	0.567	0.550
base_2016	-	-	-	-	-	0.613	0.567	0.579	0.577	0.566	0.554
base_2017	-	-	-	-	-	-	0.619	0.591	0.600	0.585	0.568
base_2018	-	-	-	-	-	-	-	0.613	0.600	0.580	0.578
base_2019	-	-	-	-	-	-	-	-	0.644	0.602	0.581
base_2020	-	-	-	-	-	-	-	-	-	0.632	0.596

Table 2: TFIDF accuracy for different base and target years using future looking corpus

6 GPT-3 embeddings

Additionally, we also use OpenAI's text embedding model [2] to obtain vectors for each K-10 filing. We use the recently released *text-ada-002* since it has a larger context window of 8192 tokens. Regardless of the number of input tokens, the model outputs an embedding vector of 1536 dimensions. Since the input context length is small for an average document in the base corpus, we only use embeddings on the future looking corpus. We truncate the model input to 8100 tokens for filings that exceed this limit. We perform the same method of training as the TF-IDF scores and use the same classification model. Results for the research ranking task are reported below:

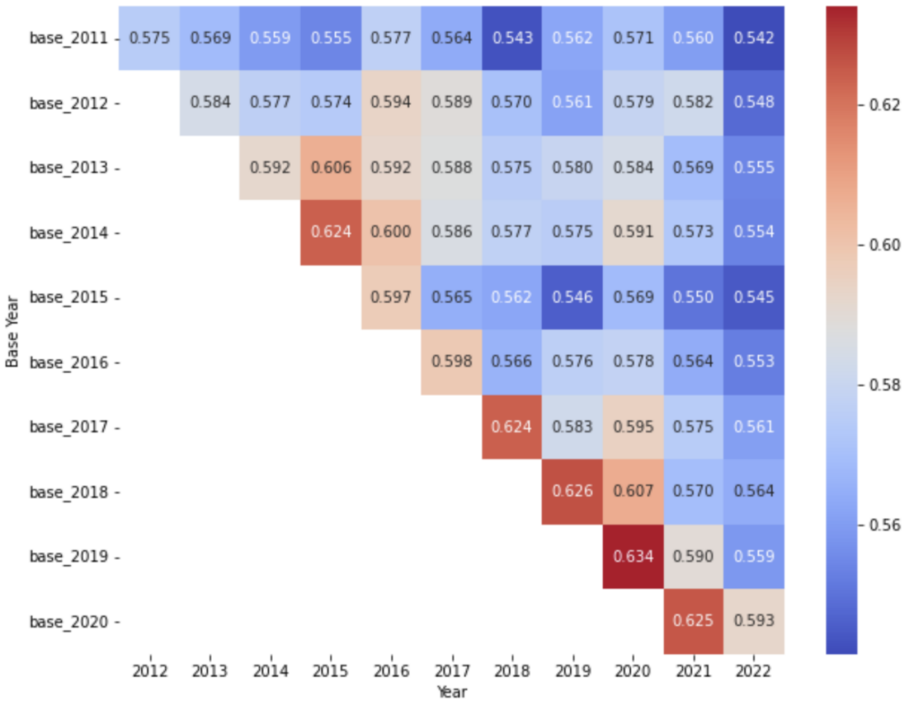

Figure 3: Heat of Accuracy GPT Base Corpus

Results for future corpus are given below:

[2] https://openai.com/blog/new-and-improved-embedding-model

Target	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
base_2011	0.575	0.569	0.559	0.555	0.577	0.564	0.543	0.562	0.571	0.560	0.542
base_2012	-	0.584	0.577	0.574	0.594	0.586	0.570	0.561	0.579	0.582	0.548
base_2013	-	-	0.592	0.606	0.592	0.588	0.575	0.580	0.584	0.569	0.555
base_2014	-	-	-	0.624	0.600	0.586	0.577	0.575	0.591	0.573	0.554
base_2015	-	-	-	-	0.597	0.565	0.562	0.546	0.569	0.550	0.545
base_2016	-	-	-	-	-	0.598	0.566	0.576	0.578	0.564	0.553
base_2017	-	-	-	-	-	-	0.624	0.583	0.595	0.575	0.561
base_2018	-	-	-	-	-	-	-	0.626	0.607	0.570	0.564
base_2019	-	-	-	-	-	-	-	-	0.634	0.590	0.559
base_2020	-	-	-	-	-	-	-	-	-	0.625	0.593

Table 3: GPT-3 embeddings accuracy for different base and target years using base corpus

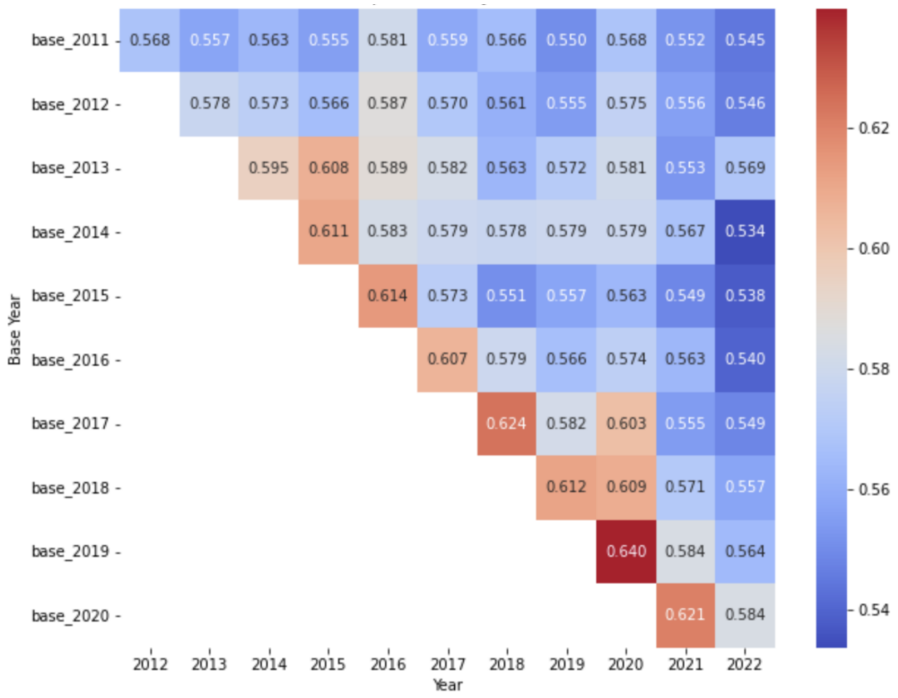

Figure 4: Heat of Accuracy GPT Future Looking Corpus

Target	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
base_2011	0.568	0.557	0.563	0.555	0.581	0.559	0.566	0.550	0.568	0.552	0.545
base_2012	-	0.578	0.573	0.566	0.587	0.570	0.561	0.555	0.575	0.556	0.546
base_2013	-	-	0.595	0.608	0.589	0.582	0.563	0.572	0.581	0.553	0.569
base_2014	-	-	-	0.611	0.583	0.579	0.578	0.579	0.579	0.567	0.534
base_2015	-	-	-	-	0.614	0.573	0.551	0.557	0.563	0.549	0.538
base_2016	-	-	-	-	-	0.607	0.579	0.566	0.574	0.563	0.540
base_2017	-	-	-	-	-	-	0.624	0.582	0.603	0.555	0.549
base_2018	-	-	-	-	-	-	-	0.612	0.609	0.571	0.557
base_2019	-	-	-	-	-	-	-	-	0.640	0.584	0.564
base_2020	-	-	-	-	-	-	-	-	-	0.621	0.584

Table 4: GPT-3 embeddings accuracy for different base and target years using future looking corpus

6.1 Fine-tuning the embeddings

Since the GPT embeddings are trained on a non-financial corpus, we experiment with a way to incorporate domain-specific information in the embeddings by fine-tuning the embeddings. Since OpenAI does not have a fine-tuning feature on the embedding model, we employ a method, inspired by contrastive learning and included in the OpenAI cookbook to customize embeddings for a task. We can learn a custom matrix through backpropagation, that when multiplied to the embedding matrix brings similar text together and different text further apart. Similar texts are defined as text that belong to the same class, while dissimilar texts belong to the opposite task.

Let $X = \{x_1, x_2 \dots x_n\}$ be a collection of texts and $Y = \{y_1, y_2 \dots y_n\}$ be the binary labels associated with the text. Let $E_x \in \mathbb{R}^d$ be an embedding of the text x that is generated from an external model and $W \in R^{d \times d}$ be a matrix we will learn to fine-tune. Take $x_{jk} = (x_j, x_k)$ where $j, k \in \{1 \dots n\}$ are randomly selected. We take $y_{jk} = 1$ if $y_k = y_j$, and $y_{jk} = -1$ otherwise. Now, to fine-tune the embeddings for the classification problem, we construct a training dataset by sampling j and k for an arbitrary N amount of times, i.e. $\mathcal{D} = \{(x_{jk}, y_{jk})\}_N$ is the fine-tuning dataset.

The learning process can then be summarized as:

$$\hat{Y}_{jk} = \text{dist}(WE_{x_j}, WE_{x_k})$$
$$W = \arg\min_W \frac{1}{N} \sum (Y_{jk} - \hat{Y}_{jk})^2$$

where dist(x, y) is the cosine similarity, $\frac{x \cdot y}{||x||||y||}$

We test the fine-tuning process on the research ranking task. To ensure a good balance, we bias the random sampling of texts to create an approximately equal amount of positive and negative text pairs. We also perform a standard test-train split and only conduct the fine-tuning process on the train set, to ensure there is no leakage in the embeddings. We experiment with multiple batch sizes, learning rates and dropout fraction to prevent overfitting and ensure optimal learning.

References

[1] Al-Ali, A. G., Phaal, R., and Sull, D. Deep learning framework for measuring the digital strategy of companies from earnings calls. In *Proceedings of the 28th International Conference on Computational Linguistics* (Barcelona, Spain

(Online), Dec. 2020), International Committee on Computational Linguistics, pp. 927–935.

[2] Araci, D. Finbert: Financial sentiment analysis with pre-trained language models. *ArXiv abs/1908.10063* (2019).

[3] Athanasakou, V., and Hussainey, K. The perceived credibility of forward-looking performance disclosures. *Accounting and Business Research 44*, 3 (2014), 227–259.

[4] Başgoze, P., and Sayin, C. The effect of r&d expenditure (investments) on firm value: Case of istanbul stock exchange. *Journal of Business Economics and Finance 2*, 3 (2013), 5–12.

[5] Chava, S., Du, W., and Paradkar, N. More than buzzwords? firms' discussions of emerging technologies in earnings conference calls. In *Firms' Discussions of Emerging Technologies in Earnings Conference Calls* (2020).

[6] Eberhart, A. C., Maxwell, W. F., and Siddique, A. R. An examination of long-term abnormal stock returns and operating performance following r&d increases. *The journal of finance 59*, 2 (2004), 623–650.

[7] Feng, F., Wang, X., He, X., Ng, R., and Chua, T.-S. Time horizon-aware modeling of financial texts for stock price prediction. In *Proceedings of the Second ACM International Conference on AI in Finance* (New York, NY, USA, 2022), ICAIF '21, Association for Computing Machinery.

[8] Glasserman, P., Krstovski, K., Laliberte, P., and Mamaysky, H. Choosing news topics to explain stock market returns. In *Proceedings of the First ACM International Conference on AI in Finance* (New York, NY, USA, 2021), ICAIF '20, Association for Computing Machinery.

[9] LI, F. The information content of forward-looking statements in corporate filings—a naïve bayesian machine learning approach. *Journal of Accounting Research 48*, 5 (2010), 1049–1102.

[10] Loukas, L., Fergadiotis, M., Androutsopoulos, I., and Malakasiotis, P. EDGAR-CORPUS: Billions of tokens make the world go round. In *Proceedings of the Third Workshop on Economics and Natural Language Processing* (Punta Cana, Dominican Republic, Nov. 2021), Association for Computational Linguistics, pp. 13–18.

[11] Ma, Z., Wang, C., Bang, G., and Liu, X. Utilization of deep learning to mine insights from earning calls for stock price movement predictions. In *Proceedings of the First ACM International Conference on AI in Finance* (New York, NY, USA, 2021), ICAIF '20, Association for Computing Machinery.

[12] Morbey, G. K. R&d: Its relationship to company performance. *Journal of Product Innovation Management: An international publication of the product development & management association 5*, 3 (1988), 191–200.

[13] Pagliaro, C., Mehta, D., Shiao, H.-T., Wang, S., and Xiong, L. Investor behavior modeling by analyzing financial advisor notes: A machine learning perspective. In *Proceedings of the Second ACM International Conference on AI in Finance* (New York, NY, USA, 2022), ICAIF '21, Association for Computing Machinery.

[14] Pataci, H., Sun, K., and Ravichandran, T. DigiCall: A benchmark for measuring the maturity of digital strategy through company earning calls. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)* (Abu Dhabi, United Arab Emirates (Hybrid), Dec. 2022), Association for Computational Linguistics, pp. 58–67.

[15] Son, G., Lee, H., Kang, N., and Hahm, M. Removing non-stationary knowledge from pre-trained language models for entity-level sentiment classification in finance, 2023.

[16] Tubbs, M. The relationship between r&d and company performance. *Research-Technology Management 50*, 6 (2007), 23–30.

[17] Tung, L. T., and Binh, Q. M. Q. The impact of r&d expenditure on firm performance in emerging markets: evidence from the vietnamese listed companies. *Asian Journal of Technology Innovation 30*, 2 (2022), 447–465.

[18] Vamvourellis, D., Toth, M., Desai, D., Mehta, D., and Pasquali, S. Learning mutual fund categorization using natural language processing. In *Proceedings of the Third ACM International Conference on AI in Finance* (New York, NY, USA, 2022), ICAIF '22, Association for Computing Machinery, p. 87–95.

[19] Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., and Mann, G. Bloomberggpt: A large language model for finance. *ArXiv abs/2303.17564* (2023).