CUBAN: Leveraging Semantic Comparables to Predict Financial Metrics Using Textual Descriptions of Companies

Siwoo Bae, *Member, IEEE*
Department of Computer Science
Vanderbilt University
Nashville, Tennessee, United States
siwoo.bae@vanderbilt.edu

David Hyde, *Senior Member, IEEE*
Department of Computer Science
Vanderbilt University
Nashville, Tennessee, United States
david.hyde.1@vanderbilt.edu

***Abstract*—Forecasting companies' financial metrics, such as profit or revenue, from textual data is typically heuristic and subjective due to the qualitative nature of business models and data. However, this paper shows that these metrics can be predicted with surprising accuracy using only a textual description of a company's business and public data from peer companies using our novel framework CUBAN (Contextual Understanding of Business performance through Analysis of Neighboring companies). We introduce a multimodal transformer model with an annotation-gating mechanism that effectively integrates textual context with financial statements. Trained on 10-K reports from public companies since 2000, our model predicts the future log revenue and operating profit from descriptions of a company's business and those of its peers (along with peer financial data), achieving a correlation coefficient (Pearson's R) of 0.78 for log revenue prediction and a 79% F1-score for operating profit classification, demonstrating its efficacy in forecasting financial performance from primarily qualitative data.**

***Index Terms*—Deep Learning, Finance, Natural Language Processing, Large Language Models, Transformers**

I. Introduction

Private equity investors and venture capitalists often face the challenge of valuing firms in the absence of quantitative data typical of public companies, relying instead on qualitative information such as textual descriptions of business models, pitch decks, and technical reports [1]. Even in the public markets, where greater quantitative data is typically available, equity traders frequently seek further insights from textual data to derive advantages [2]. However, deriving accurate assessments from qualitative data has proven to be a difficult task. For example, despite ample qualitative information that may exist about a startup, such information is unlikely to accurately assess product-market fit, the lack of which is one of the leading causes of startup failure [3].

Recent advances in natural language processing (NLP) have introduced various useful techniques for encoding text as numerical representations, such as SBERT [4]. Thus, one can apply methodologies for numerical data on qualitative and textual data. While various recent works have leveraged learning and NLP approaches on text for financial forecasting (e.g., [5]–[7]), the present paper offers a novel method that is distinct in its architecture, source data, and outputs. Our proposed deep learning approach estimates a firm's future revenue and profitability based solely on textual descriptions of its business and publicly available data from comparable companies. Inspired by comparable company analysis [8], [9], our framework, CUBAN (Contextual Understanding of Business performance through Analysis of Neighboring companies), leverages information from semantically similar firms to make predictions. To the best of our knowledge, this is the first large-scale, multimodal approach aimed at predicting a firm's future performance where the only information about that firm used at query time is its textual business description.

We note that with the types of data considered in this paper, there are two particularly relevant challenges:

- **Concept drift**: Market dynamics evolve over time, altering consumer behavior and rendering earlier data less relevant. Learning models often assume that data distributions are independent and identically distributed (i.i.d.) over time [10], which leads to performance degradation when market conditions change.
- **Multimodal heterogeneity**: In business and finance, information is often expressed through both text and structured (i.e., tabular, quantitative) data. The ability to combine these modalities effectively is essential, as relying on just one may lead to misleading conclusions. However, without normalization, integrating these diverse data types can result in numerical instability and model convergence issues.

In response, CUBAN addresses these challenges by integrating a pretrained text embedding model, a similarity search algorithm, and a machine learning model designed to predict revenue and profitability using multimodal data from textual descriptions and financial statements.

As demonstrated in Figure 2, simply performing a similarity search alone does not provide adequate information, as the search not only identifies peers but also other connections, including customer/supplier relationships and companies with complementary products. This becomes problematic for a

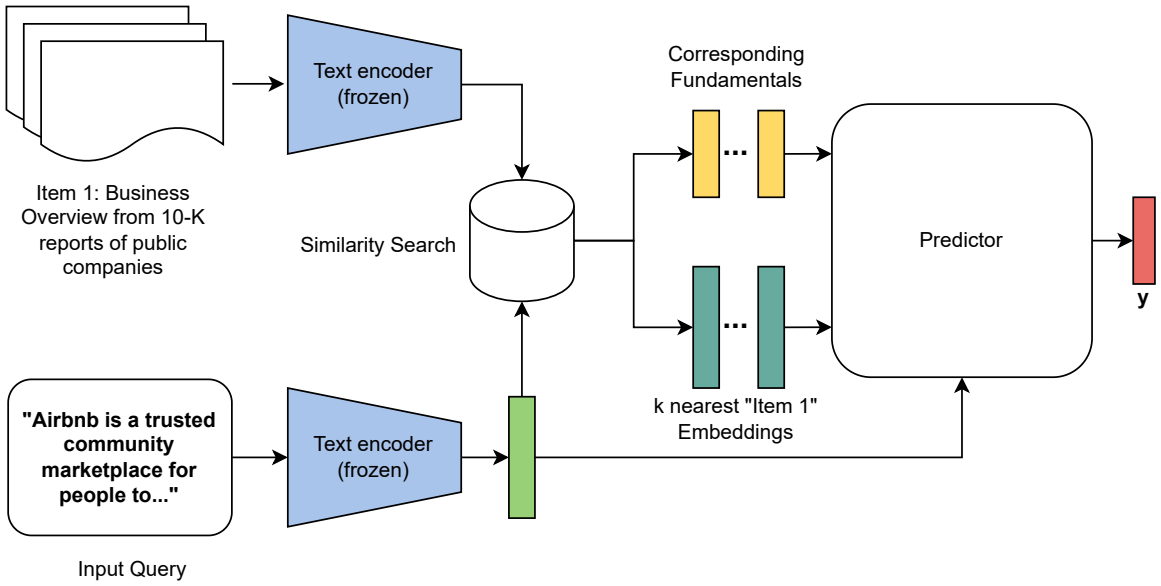


Fig. 1: Overview of our CUBAN framework. A company description is encoded into a vector of length 1024 using a text embedding model. Similarity search finds similar companies in this embedding space. The financial data of these companies are passed into a machine learning model, which ultimately predicts financial metrics for the target company.

naive semantic search engine that aims to identify peer firms exclusively. However, we note that this limitation can be turned into an advantage by integrating an additional analyst model that addresses the complexity of analyzing the dynamics among these interconnected firms.

Our results show that with CUBAN, mere textual descriptions of a company's business and the businesses of its peers (along with peer financial data) can predict metrics like revenue and profitability with surprising accuracy.

In summary, we have made three key contributions:

- We introduce CUBAN, a novel framework that leverages information from semantically similar firms to improve the accuracy of financial forecasting.
- We develop a novel multimodal transformer model, LCCT, that integrates textual descriptions and structured financial data using an annotation-gating mechanism.
- We empirically validate our model on 10-K reports from public companies, demonstrating its efficacy in predicting financial performance with a correlation coefficient of 0.78 for log revenue prediction and a 79% F1-score for operating profit classification.

II. Related Work

A. Machine Learning in Finance

The application of machine learning (ML) techniques in finance has gained significant momentum, offering new approaches to tackle complex financial problems. Traditionally, financial forecasting and analysis have relied heavily on structured financial data such as balance sheets, income statements, and cash flows. While these quantitative metrics are essential for evaluating a company's financial health, they often fail to capture qualitative aspects, particularly for early-stage firms with limited historical data. Advancements in ML have enabled the incorporation of alternative data sources, including textual information from company reports, news articles, and social media, to enhance predictive models. [11] provided a comprehensive review of recent ML applications in finance, highlighting its potential to improve forecasting accuracy and risk assessment. They emphasized that ML methods could address complex, nonlinear relationships in financial data, which traditional econometric models might miss. In the context of asset pricing, [12] demonstrated how ML techniques could be employed to better predict asset returns by capturing intricate patterns in large datasets. Their work showed that models like random forests and neural networks outperformed traditional linear models in out-of-sample predictions. For credit risk assessment, [13] proposed integrating unsupervised and supervised ML algorithms, showing that such hybrid models outperform traditional methods. Similarly, [14] utilized neural network rule extraction and decision tables to evaluate credit risk, providing insights into the reasoning behind ML decisions. Detecting fraudulent activities is another area where ML has made significant contributions. [15] developed a machine learning approach to detect accounting fraud in publicly traded U.S. firms, emphasizing the effectiveness of ML in identifying anomalies that may not be evident through conventional analysis.

B. Text Embedding

Text embedding techniques are essential for converting textual data into numerical representations that ML models can process. Early methods like Term Frequency-Inverse Document Frequency (TF-IDF) [16] and Word2Vec [17] focused on

capturing word-level semantics based on word frequencies and co-occurrences. Word2Vec, for instance, uses neural networks to produce word embeddings that reflect semantic similarities between words.

The introduction of the transformer architecture [18] revolutionized text representation by enabling models to capture contextual relationships between words in a sentence more effectively. Building on this architecture, Bidirectional Encoder Representations from Transformers (BERT) [19] allowed for bidirectional training, understanding the context of a word based on both its left and right surroundings. Sentence-BERT (SBERT) [4] further advanced this field by generating sentence-level embeddings using a siamese network architecture. SBERT enables the comparison of semantically meaningful sentence embeddings using cosine similarity, significantly reducing computational overhead and improving performance on tasks like semantic textual similarity and clustering.

In the financial domain, these advanced text embedding techniques have been utilized to analyze large volumes of unstructured text data. For instance, [20] developed a knowledge-driven text-embedding approach to analyze firm reports for volatility prediction, demonstrating the effectiveness of combining domain-specific knowledge with advanced embedding techniques. [21] used Word2Vec and SBERT to perform clustering of financial institutions based on annual report data.

C. Company Similarity Analysis

Identifying similar firms or peer companies is crucial in various financial applications, including valuation, risk assessment, and investment strategy development. Traditionally, industry classifications such as the Global Industry Classification Standard (GICS) have been used to group companies. However, these classifications can be coarse-grained and may not capture nuanced differences, especially in rapidly evolving industries [22].

Recent research has explored the use of Natural Language Processing (NLP) and machine learning techniques to derive more granular measures of company similarity based on textual data. For instance, [23] introduced the concept of text-based network industries, grouping firms based on the similarity of product descriptions in their SEC filings. This approach allows for dynamic industry groupings that better reflect the competitive landscape. [24] utilized Word2Vec to create embeddings of news articles, allowing the identification of peer firms based on semantic content, and in a similar vein, [25] proposed defining peer firms using common analyst coverage, arguing that analysts' choices provide insight into firm relatedness beyond traditional classifications. Their method resulted in more homogeneous groups compared to standard industry classifications. Leveraging large language models (LLMs), [26] embedded business descriptions from SEC filings to reproduce GICS classifications and reveal similarities in financial performance metrics. Their findings suggest that LLMs can effectively capture the semantic content of business descriptions, enabling more nuanced company similarity analysis. Furthermore, [27] developed an industry peer grouping system based on artificial intelligence that clusters companies using machine learning algorithms on various attributes, including textual data from financial reports. Their system demonstrated better performance over traditional classification schemes.

Lastly, we highlight [28], which conducted a comparative study on measuring company similarity using financial statements. The results illustrated the importance of combining structured financial data with textual analysis. They emphasized that models based on graph theory and interconnected structured data could enhance the accuracy of similarity assessments.

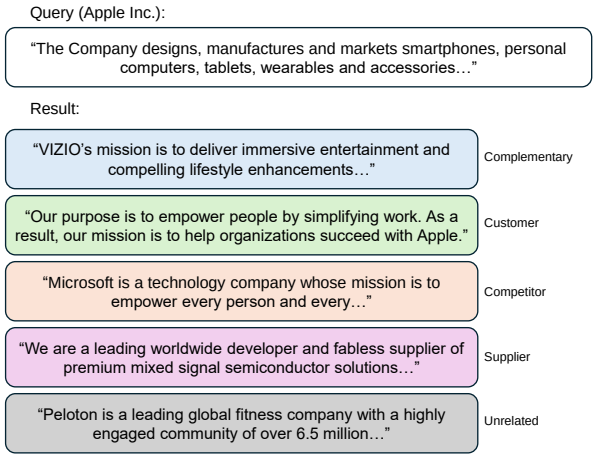


Fig. 2: Identifying the top k nearest neighbors to Apple's Item 1 text, as detailed in the 2023 10-K report, produces not only competitors but also identifies customer/supplier relationships and companies offering complementary products.

III. Data Collection and Preprocessing

A. Data Collection

The United States Securities and Exchange Commission (SEC) requires publicly traded companies to file 10-K annual reports, which provide a comprehensive overview of a company's financial condition, including detailed information on business activities, fiscal performance, and governance. For this study, we collected 193,000 10-K reports from more than 30,000 companies dating back to 2000 using the SEC EDGAR API. We specifically extracted and processed the "Item 1. Business Overview" section and used the Refinitiv Eikon API to obtain standardized annual and quarterly financial statements corresponding to the relevant fiscal years and companies. The dataset was then divided into three segments: 2000–2017 for training, 2018–2020 for validation, and 2021–2024 for testing. A one-year gap between these segments was maintained to prevent data leakage.

B. Data Preprocessing and Feature Engineering

Each organization's Item 1 text was transformed into a 1,024-dimensional embedding vector. The financial data, comprising balance sheets, income statements, and cash flow

statements, formed the basis of our raw financial data. Since standard accounting procedures typically omit entries with values of 0, all NaN entries have been replaced with 0. The target variables are generated by taking log revenue and operating profit margin values of the next year (or quarter).

The feature engineering process involves concatenating all of the following transformations to the raw financial data:

- symmetric logarithm $symlog(x) = \text{sgn}(x)\log(|x|+1)$,
- dividing by total assets,
- dividing by revenue,
- dividing by market capitalization, and
- symmetric logarithm of the yearly (or quarterly) growth.

Ultimately, the embedding vectors are described by the dimension d_e, and the fundamental features have a dimension of d_f, with values $d_e = 1024$ and $d_f = 224$.

The symmetric logarithm transformation was chosen to address the fat-tailed distributions characteristic of financial data, which often include both positive and negative values while being invertible and less dependent on the data distribution compared to traditional standardization methods like z-score normalization.

Finally, the data are organized into two comprehensive datasets: the annual and the quarterly versions. The annual dataset is compiled by performing an inner join on all three data categories: embeddings, fundamentals, and targets. On the other hand, the quarterly dataset uses a time-sensitive join that aligns quarterly fundamentals with embeddings within a year. Because the targets derive from the fundamentals, they share the same time intervals, allowing for a simple inner join. This process generates a yearly dataset with 56,844 training samples, 4,297 validation samples, and 7,695 test samples. Similarly, the quarterly dataset includes 202,415 training samples, 15,199 validation samples, and 7,746 test samples.

C. Training and Validation

To simulate the prediction of a financial metric, a random sample (referred to as the query) is selected from the dataset. Subsequently, we collect its k nearest neighbors while excluding the query's fundamentals. Thus, each training batch for the predictor model contains:

- **Query Embeddings:** The "Item 1" business description embeddings for the randomly sampled query company, with dimensions (B, d_e), where d_e is the dimensionality of the encoded textual embeddings from the pretrained model.
- **Neighbor Embeddings and Financial Data:** For each query company, we retrieve the k nearest neighbors based on cosine similarity of their business description embeddings. These neighbor embeddings are combined with their corresponding financial data. The resulting input for the neighbors has dimensions $(B, k, d_e + d_f)$, where k is the number of neighbors.
- **Time Differences:** The differences in years between the filing dates of the query company and those of its neighboring companies are incorporated as extra features, structured with dimensions (B, k). This is similar to the role of positional encoding in conventional transformer architectures.
- **Target Variable:** For each query company, the target variable (such as next year's revenue or operating profit margin) is denoted by $(B, 1)$.

At each epoch, the model is evaluated on the validation set. The model that performs best on the validation set is then used for evaluation on the test set to ensure generalization performance. In addition, the validation and test set are constructed so that the model is evaluated only by query companies that *do not exist in the training set* to make the testing condition more rigorous. In other words, the model does not face any query companies that appear in the validation or test sets during training, whether as a query or a neighbor.

IV. Methods

A. Problem Formulation

The primary objective of this research is to develop a predictive model capable of estimating a query company's future financial metrics, such as revenue and profitability, using only its textual business description and the publicly available financial data of comparable companies.

Formally, let $\mathcal{C} = c_1, c_2, \ldots, c_N$ be a set of companies with known data. Each company c_i has an associated textual business description $\boldsymbol{e}_i$ and financial data $\boldsymbol{F}_i$, which may include balance sheets, income statements, and cash flow statements. Also, let c_q be the query company with a textual business description $\boldsymbol{e}_q$, for which we aim to predict future financial metrics y_q (e.g., the revenue or profitability ratios for the next period).

Our goal is to learn a function f that maps the description of the query company and the data from similar companies to an estimate of its future financial metrics:

$$\hat{y}_q = f(\boldsymbol{e}_q, \{(\boldsymbol{e}_i, \boldsymbol{F}_i)\}_{i \in \mathcal{I}_k}), \tag{1}$$

where $\hat{y}_q$ is an estimate of y_q and $\mathcal{I}_k$ is the index set of the top k companies most similar to c_q based on the similarity of their business descriptions.

In seeking such an f, we make the following assumptions:

- Companies with similar business descriptions are likely to contain information beneficial for predicting the query company's performance metrics as they are exposed to similar market dynamics.
- The text embedding model effectively captures the semantic meaning of business descriptions, enabling accurate similarity assessments.
- The financial data of comparable companies are timely and relevant, reflecting current market conditions applicable to the query company.

B. CUBAN

Our CUBAN framework (see Figure 1) consists of three main components: a pretrained text embedding model, a similarity search algorithm, and the predictor model, which fuses

textual embeddings and financial data for financial prediction. These components work as follows.

1) Long Range Text Embedding: We begin by encoding the textual description of a company's business model using mGTE [29], a state-of-the-art text embedding model. This model transforms the input text into a 1,024-dimensional vector, capturing the semantic nuances of the business description. We opted for this embedding model because it has a long enough context window to cover most company descriptions in our dataset, while still being relatively computationally lightweight. mGTE is capable of processing English text inputs up to a maximum of 8,192 tokens, which equates to roughly 7,000 words or 30,000 characters. In contrast, most company descriptions are typically less than 6,000 tokens.

2) Similarity Search: Once the business model is embedded, we perform a similarity search to identify the top k companies with the most similar embeddings within a year based on cosine similarity. Figure 2 shows an example result of our similarity search conducted for Apple. For this search, we only consider peer companies whose financial data is less than one year old (excluding, e.g., firms that have gone out of business or gone private) in order to improve the accuracy of our predictions. The financial data of these peer companies, including their most recent financial statements, are then retrieved for further analysis.

3) Financial Metric Prediction: Ultimately, all the inputs are provided to a supervised machine learning model designed to forecast desired financial metrics for the upcoming year or quarter. Through empirical studies, we show that choosing various machine learning models can lead to diverse results depending on other hyperparameters such as k. Section V provides an in-depth analysis to determine the optimal value of k and the machine learning model that achieves superior performance.

C. Prediction Models

In our experiments, we assess the predictive power of several different machine learning models.

a) MeanPool+Linear and MeanPool+XGBoost: These models provide a baseline that can be used within CUBAN's framework. The central concept is that the query company exhibits market behaviors and asset patterns analogous to those of comparable firms. Accordingly, the input consists of (1) the query company's embedding and (2) the averaged fundamental data from the k nearest neighbor firms. Linear and logistic regressions, both incorporating L^2 regularization, are used. Additionally, we test XGBoost [30] on this dataset to account for non-linear dynamics.

b) MLP+MeanPool: This model first integrates query and embedding data using a gating mechanism, processes the result through a 3-layer element-wise MLP with ReLU activations, and finally performs mean pooling before the prediction output. The gating mechanism fuses the neighbor embeddings N with the corresponding tabular financial data $T \in \mathbb{R}^{k \times d_f}$. Each neighbor embedding gates its corresponding row of financial data using the CrossGLU module. The gating operation is defined as:

$$\text{GatedOutput} = \sigma(W_g \cdot E) \odot (W_f \cdot F), \tag{2}$$

where:

- $E \in \mathbb{R}^{k \times d_e}$ represents the neighbor embeddings.
- $F \in \mathbb{R}^{k \times d_f}$ represents the tabular financial data for the neighbors.
- $W_g \in \mathbb{R}^{d_e \times d_{\text{output}}}$ is the learned weight matrix for the gating signal.
- $W_f \in \mathbb{R}^{d_f \times d_{\text{output}}}$ is the learned weight matrix for transforming the financial data.
- σ is the sigmoid activation function, and $\odot$ denotes element-wise multiplication.

The gating mechanism allows each neighbor embedding to modulate the corresponding financial data before it is passed to subsequent layers. This ensures that the textual context of each neighbor directly influences how its financial data is used in the model.

c) LCCT: Our proposed Locally Contextualized Comparative Transformer (LCCT) model (Figure 3) is a novel transformer-based model that integrates textual embeddings and financial data, allowing for the joint processing of business descriptions and structured financial data. The architecture is comprised of several key components, including cross-attention, self-attention, a feed-forward network, and a gating mechanism for multimodal fusion. Below, we outline each layer in detail.

The cross-attention layer attends to the query company's textual embedding and retrieves relevant information from the financial data of similar companies. Given a query embedding $Q \in \mathbb{R}^{1 \times d_e}$ and a set of neighbor embeddings $N \in \mathbb{R}^{k \times d_e}$, the query is first broadcast to match the shape of N, resulting in $Q' \in \mathbb{R}^{k \times d_e}$. The cross-attention operation is then applied as follows:

$$A_{\text{cross}} = \text{MultiHeadAttn}(Q', E, E) \in \mathbb{R}^{k \times d_e}.$$

This attention mechanism allows the model to attend to and aggregate relevant financial data from similar companies.

After cross-attention, the self-attention mechanism refines the representation by attending to internal relationships among the neighbor embeddings E:

$$A_{\text{self}} = \text{MultiHeadAttn}(E, E, E) \in \mathbb{R}^{k \times d_e}.$$

The self-attention mechanism ensures that the model captures the interdependencies between neighbors.

Subsequently, each attention output is passed through a position-wise feed-forward network to further transform the representations:

$$\text{FFN}(x) = W_2 \cdot \text{Activation}(W_1 \cdot x + b_1) + b_2,$$

where $W_1 \in \mathbb{R}^{d_e \times d_{\text{ff}}}$ and $W_2 \in \mathbb{R}^{d_{\text{ff}} \times d_e}$ are learned weight matrices, and d_{ff} is the hidden dimension of the feed-forward network. The activation function can vary (e.g., ReLU, GELU,

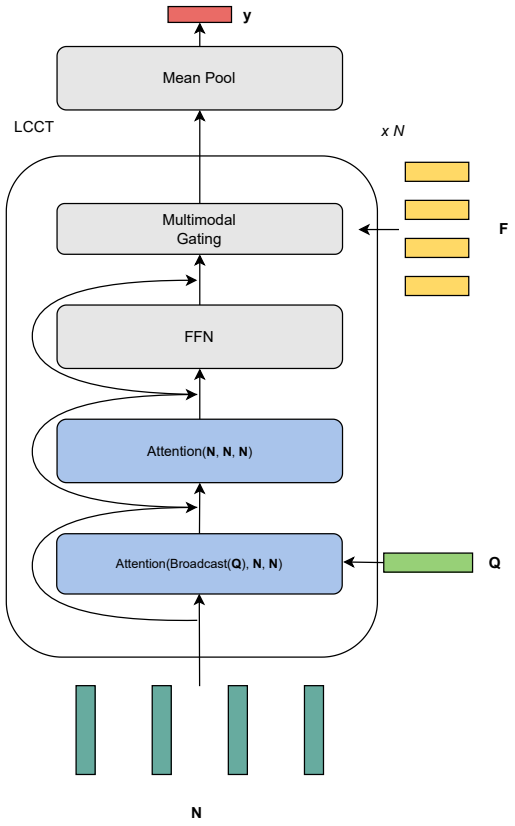


Fig. 3: Schematic of the LCCT model architecture and its use. An LCCT takes in N textual embedding vectors of company descriptions, along with N associated vectors F corresponding to those companies' financial data. Along with a description Q of a target company, this data is passed through several attention, fusion, and pooling units to ultimately predict the financial metrics for the target company.

or SwiGLU), but we have empirically found that vanilla ReLU activation performed best on our validation sets.

Finally, the information in the fundamentals data is fused using the CrossGLU gating mechanism (see Eq. 2), and layer normalization [31] is applied after each of the attention blocks and the feed-forward block to stabilize training. LCCT also employs residual connections help retain the original input information, preventing gradient vanishing during backpropagation, as illustrated in Figure 3. For all the experiments carried out in this study, we utilized $N = 4$, meaning there were four LCCT layers stacked sequentially to produce the final output.

V. Experiments

We evaluated several machine learning models across two primary tasks: predicting revenue and profitability for the subsequent period, using both annual and quarterly datasets. Revenue prediction was framed as a regression task to estimate the logarithm of the expected revenue, while profitability prediction involved classifying whether the operating profit would be positive or negative.

TABLE I: Performance evaluation on the annual dataset

	Log Revenue Regression (Pearson's R)				
Model	k=4	k=8	k=16	k=32	k=64
MeanPool+Linear	0.019	-0.020	-0.02	0	0
MeanPool+XGBoost	0.726	0.726	0.720	0.713	0.712
MLP+MeanPool	**0.759**	**0.774**	0.773	**0.778**	0.783
LCCT	0.752	0.768	**0.778**	0.776	**0.791**
	Profitability Classification (F1)				
Model	k=4	k=8	k=16	k=32	k=64
MeanPool+Logistic	0.748	0.747	0.753	0.750	0.752
MeanPool+XGBoost	0.777	0.782	0.783	0.786	0.783
MLP+MeanPool	0.792	0.784	0.779	0.785	**0.806**
LCCT	**0.793**	**0.795**	**0.796**	**0.797**	0.793

TABLE II: Performance evaluation on the quarterly dataset

	Log Revenue Regression (Pearson's R)				
Model	k=4	k=8	k=16	k=32	k=64
MeanPool+Linear	0	0	0	0	0
MeanPool+XGBoost	0.631	0.628	0.635	0.653	0.635
MLP+MeanPool	**0.695**	**0.686**	**0.703**	**0.702**	**0.720**
LCCT	0.663	0.661	0.698	0.693	0.712
	Profitability Classification (F1)				
Model	k=4	k=8	k=16	k=32	k=64
MeanPool+Logistic	0.695	0.669	0.739	0.742	0.705
MeanPool+XGBoost	0.770	0.771	0.777	0.781	**0.780**
MLP+MeanPool	0.768	0.770	**0.779**	0.771	0.759
LCCT	**0.778**	**0.778**	**0.779**	**0.782**	**0.780**

For revenue prediction, we used Pearson's correlation coefficient (R) as it captures the linear relationship between the predicted and actual logarithmic revenues, providing a robust and intuitive measure of model efficacy in identifying revenue trends rather than absolute values. This is particularly important given the nature of the problem, where understanding the relative rankings between predicted values among firms matters more than precise point predictions. To evaluate profitability classification, we used the F1-score. Due to variations in label distribution over time arising from fluctuating market conditions, and considering that both false positives and false negatives carry considerable implications here, the F1-score provides an appropriate assessment of the model's classification performance.

In our comparisons, the models evaluated comprise baseline methods such as MeanPool+Linear and MeanPool+XGBoost, as well as our innovative models, including MLP+MeanPool and LCCT. The results, presented in Tables I and II, emphasize the durability of the LCCT model, which consistently exceeded baseline methods, especially in revenue regression with Pearson's R of 0.791 and profitability classification with an F1-score of 0.797.

While the MLP+MeanPool model showed better results for revenue prediction in the quarterly dataset, LCCT demonstrated strong performance across most scenarios, indicating that the attention mechanism serves as an effective inductive bias. Performance discrepancies between models suggest that the optimal architecture may depend on factors such as time

TABLE III: Data Ablation Analysis on Log Revenue Regression on the Annual Dataset (Pearson's R) with $k = 32$

Model	LCCT	Linear	XGBoost
query+embed+fund	**0.776**	0	0.713
query+embed	0.743	**0.757**	0.710
query+fund	**0.776**	0	0.714
embed+fund	**0.639**	0.002	0.632

resolution, number of neighbors, and prediction target.

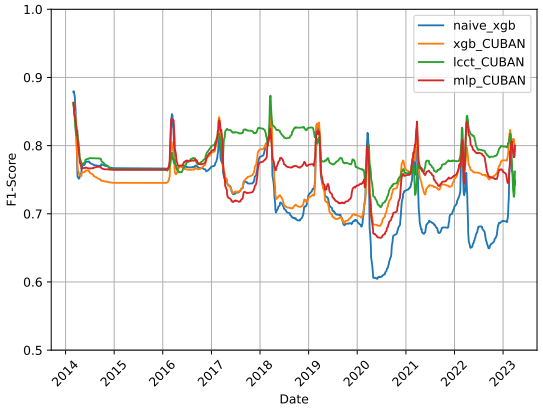
Conducting additional ablation studies (Table III) has emphasized the essential function of multimodal integration in our framework's performance. The table illustrates the predictive power of models based on LCCT, MeanPool+Linear, and MeanPool+XGBoost when various pieces of data are removed (the query company's embedding, nearest neighbor embeddings, or nearest neighbor fundamentals). The table shows that excluding the query embedding resulted in a more pronounced reduction in performance compared to omitting financial data, highlighting the significance of textual descriptions in forecasting a company's future financial metrics.

Additionally, we demonstrate that CUBAN with LCCT maintains robustness against concept drift, as anticipated. Figure 4 compares a naïve approach, which involves training XGBoost only on query embeddings without comparable data, against CUBAN-based methods. The result suggests that CUBAN with LCCT experiences the slowest decline in performance over time. Additionally, in the operating profit classification task, LCCT is the only model that does not experience seasonal performance drops, which implies that our approach is relatively robust against changing market dynamics. We also note that LCCT has both a higher maximum and higher minimum scores than all other baselines.

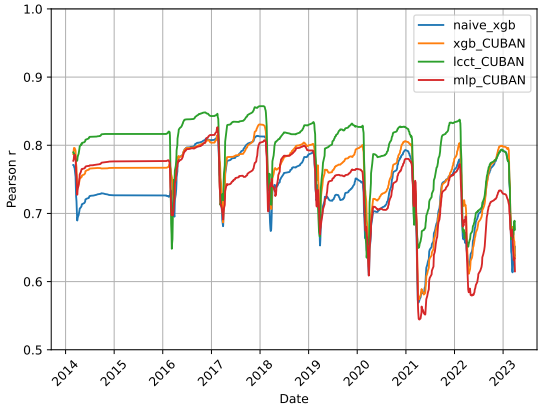
VI. Conclusions and Future Work

We presented CUBAN, a novel framework that combines qualitative and quantitative data to forecast a company's financial performance. By leveraging advanced text embedding techniques alongside financial data from comparable firms, we demonstrate that qualitative business descriptions can serve as a powerful predictor of future performance. Our findings, which include a Pearson's R of 0.79 for predicting revenue and an F1-score of 80% for classifying profitability, underscore the benefits of merging NLP techniques with financial modeling. The results reveal that textual business descriptions, when properly integrated with financial data from comparable firms, offer substantial predictive power. This aligns with the growing interest in applying natural language processing to domains traditionally reliant on structured data, such as finance.

The robust performance of our novel LCCT model across various tasks highlights the necessity of capturing semantic associations in textual data. Additionally, the capability to amalgamate data from similar companies enhances its predictive power by resembling traditional company peer analysis. Nonetheless, the differing performance across models and datasets implies that no one model is universally superior, suggesting value in selecting models tailored to specific needs.



(a) Evolution of F1-score over time on the annual test set with $k = 32$.



(b) Evolution of R over time on the annual test set with $k = 32$.

Fig. 4: Evolution of evaluation metrics over time on the annual set trained on 2000-2013 and tested on 2014-2023. (a) Operating profit classification. (b) Log revenue regression.

Note that neural networks tend to scale well as the number of neighbors k increases, despite the common perception that traditional machine learning models tend to perform much better on structured, tabular datasets [32], [33]. We attribute the success of our approach to the rigorous data preprocessing and keeping the information flow for tabular dataset rather shallow.

A notable finding from our ablation study (Table III) is the critical role of textual data in making predictions. The model's greater reliance on business descriptions than on financial fundamentals indicates the substantial value that qualitative data provides in assessing a company's prospects. Nevertheless, this raises questions about the selection of comparable companies. While our model uses cosine similarity on textual embeddings, the definition of similarity may vary between the model and human analysts, suggesting the need for further exploration of hybrid approaches that incorporate both algorithmic methods and expert judgment.

In finance, CUBAN provides a valuable tool for situations where data is sparse or unavailable, such as with early-

stage companies or in industries with limited transparency. By relying on textual business descriptions, the model offers an alternative method for financial evaluation, potentially supporting decisions in investment, M&A, and strategic planning. Furthermore, the ability to extend this framework to other sectors or geographic regions underscores its versatility.

More broadly, this work contributes to the field of multimodal learning by demonstrating a successful integration of qualitative and quantitative data for financial prediction. The combination of text and structured data allows for a more comprehensive analysis of business models, particularly in cases where financial metrics alone may be insufficient. Beyond finance, the flexibility of the CUBAN framework suggests potential applications in other fields where qualitative data plays a central role, such as law or healthcare.

Despite these promising results, several limitations remain. Reliance on high-quality textual data could pose challenges, especially in regions where such data is less standardized or unavailable. For instance, since Item 1 within 10-K reports is generally formal and comprehensive, examining a private business or start-up might necessitate an extra measure of creating in-depth reports in natural language to align with the style and depth found in Business Overview sections for reliable forecasts using CUBAN. Additionally, the complexity of transformer-based models can hinder interpretability, which is crucial in fields like finance where stakeholders require clear justifications for decisions. Finally, while our framework addresses historical data, adapting to rapidly evolving market conditions remains an area for further development.

In the future, research might aim to improve the model's interpretability by integrating explainability methods like attention visualization or model distillation. Extending the model to cover more data sources, including social media sentiment or industry reports, could enhance its stability. Combining expert knowledge with algorithmic insights could provide a route to more precise predictions, especially in unstable or emerging markets.

REFERENCES

[1] P. A. Gompers, W. Gornall, S. N. Kaplan, and I. A. Strebulaev, "How do venture capitalists make decisions?," *Journal of Financial Economics*, vol. 135, no. 1, pp. 169–190, 2020.

[2] H. Lee, M. Surdeanu, B. MacCartney, and D. Jurafsky, "On the importance of text analysis for stock price prediction.," in *LREC*, vol. 2014, pp. 1170–1175, 2014.

[3] M. Cantamessa, V. Gatteschi, G. Perboli, and M. Rosano, "Startups' roads to failure," *Sustainability*, vol. 10, no. 7, p. 2346, 2018.

[4] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," Aug. 2019.

[5] T. Swathi, N. Kasiviswanath, and A. A. Rao, "An optimal deep learning-based lstm for stock price prediction using twitter sentiment analysis," *Applied Intelligence*, vol. 52, no. 12, pp. 13675–13688, 2022.

[6] Y. Li and Y. Pan, "A novel ensemble deep learning model for stock prediction based on stock prices and news," *International Journal of Data Science and Analytics*, vol. 13, no. 2, pp. 139–149, 2022.

[7] F. Z. Xing, E. Cambria, and R. E. Welsch, "Natural language based financial forecasting: a survey," *Artificial Intelligence Review*, vol. 50, no. 1, pp. 49–73, 2018.

[8] R. G. Bowman and S. R. Bush, "Using comparable companies to estimate the betas of private companies," *Journal of Applied Finance, Forthcoming*, 2007.

[9] B. Hottenhuis, "Investigating an alternative approach to saas company valuation: using 'rule of 40' metrics as indicators of enterprise value," Master's thesis, University of Twente, 2020.

[10] M. Dundar, B. Krishnapuram, J. Bi, and R. B. Rao, "Learning classifiers when the training data is not iid," in *IJCAI*, vol. 2007, pp. 756–61, Citeseer, 2007.

[11] D. Hoang and K. Wiegratz, "Machine learning methods in finance: Recent applications and prospects," *European Financial Management*, vol. 29, pp. 1657–1701, Nov. 2023.

[12] S. Gu, B. Kelly, and D. Xiu, "Empirical Asset Pricing via Machine Learning," *The Review of Financial Studies*, vol. 33, pp. 2223–2273, May 2020.

[13] W. Bao, N. Lianju, and K. Yue, "Integration of unsupervised and supervised machine learning algorithms for credit risk assessment," *Expert Systems with Applications*, vol. 128, pp. 301–315, Aug. 2019.

[14] B. Baesens, R. Setiono, C. Mues, and J. Vanthienen, "Using Neural Network Rule Extraction and Decision Tables for Credit-Risk Evaluation," *Management Science*, vol. 49, pp. 312–329, Mar. 2003.

[15] Y. Bao, B. Ke, B. Li, Y. J. Yu, and J. Zhang, "Detecting Accounting Fraud in Publicly Traded U.S. Firms Using a Machine Learning Approach," *Journal of Accounting Research*, vol. 58, pp. 199–235, Mar. 2020.

[16] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[17] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Å. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, (Red Hook, NY, USA), pp. 6000–6010, Curran Associates Inc., 2017. event-place: Long Beach, California, USA.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (J. Burstein, C. Doran, and T. Solorio, eds.), (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.

[20] Y. Yang, K. Zhang, and Y. Fan, "Analyzing Firm Reports for Volatility Prediction: A Knowledge-Driven Text-Embedding Approach," *INFORMS Journal on Computing*, vol. 34, pp. 522–540, Jan. 2022.

[21] Z. Ran, "Clustering financial institutions with soft information: A computational linguistics approach," *Available at SSRN 4954166*, 2024.

[22] S. Bhojraj, C. M. C. Lee, and D. K. Oler, "What's My Line? A Comparison of Industry Classification Schemes for Capital Market Research," *Journal of Accounting Research*, vol. 41, pp. 745–774, Dec. 2003.

[23] G. Hoberg and G. Phillips, "Text-Based Network Industries and Endogenous Product Differentiation," *Journal of Political Economy*, vol. 124, pp. 1423–1465, Oct. 2016.

[24] T. Kee, "Peer Firm Identification Using Word Embeddings," in *2019 IEEE International Conference on Big Data (Big Data)*, pp. 5536–5543, 2019.

[25] M. Kaustia and V. Rantala, "Common Analysts: Method for Defining Peer Firms," *Journal of Financial and Quantitative Analysis*, vol. 56, pp. 1505–1536, Aug. 2021.

[26] D. Vamvourellis, M. Toth, S. Bhagat, D. Desai, D. Mehta, and S. Pasquali, "Company Similarity using Large Language Models," 2023.

[27] G. Bonne, A. W. Lo, A. Prabhakaran, K. W. Siah, M. Singh, X. Wang, P. Zangari, and H. Zhang, "An Artificial Intelligence-Based Industry Peer Grouping System," *The Journal of Financial Data Science*, vol. 4, pp. 9–36, Apr. 2022.

[28] Z. Zhang, "A Comparative Study on Measuring Similarity between Companies Based on Financial Statements," *Advances in Economics and Management Research*, vol. 8, p. 147, Nov. 2023.

[29] X. Zhang, Y. Zhang, D. Long, W. Xie, Z. Dai, J. Tang, H. Lin, B. Yang, P. Xie, F. Huang, M. Zhang, W. Li, and M. Zhang, "mGTE: Generalized Long-Context Text Representation and Reranking Models for Multilingual Text Retrieval," July 2024.

[30] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference*

on Knowledge Discovery and Data Mining, (San Francisco California USA), pp. 785–794, ACM, Aug. 2016.

[31] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016.

[32] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Information Fusion*, vol. 81, pp. 84–90, 2022.

[33] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?," in *Advances in Neural Information Processing Systems* (S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds.), vol. 35, pp. 507–520, Curran Associates, Inc., 2022.