Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A BERT-Based Framework for Extracting Business Insights from Financial Reports

Prasant Kumar Rout

prasant.rt@gmail.com

Sri Sri University, Cuttack, Odisha

Binita Nanda

Sri Sri University, Cuttack, Odisha

Sunil Kumar Dhal

Sri Sri University, Cuttack, Odisha

Research Article

Keywords: BERT, Financial Report Analysis, NLP, Sentiment Analysis, Risk Factor Extraction, FinBERT, Business Intelligence, Transformer Models, Financial Text Analytics

Posted Date: March 31st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9209699/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Financial reports bear a lot of textual information, which is normally unstructured and imperative to business decision-making. Nevertheless, it is difficult to draw meaningful conclusions out of such huge and sophisticated papers through the conventional methods of analysis. This paper is an attempt to provide a BERT-based system to extract business insights in financial reports in an automated manner. Using the ability of Bidirectional Encoder Representations of Transformer to contextualize textual information (annual reports, management discussions, financial statements, etc.), the model proposed then analyzes the textual information to extract important information (sentiment, risk factors, performance indications, etc.).

The framework includes the data preprocessing, domain-specific weighting of a pre-trained BERT, and usage of classification and information extracting methods. The research experimental results prove that BERT-based system is much more operative than the old-style machine learning (ML) models in terms of accuracy, precision and F1-score. Moreover, the inferences that were to be made are of good assistance to all stakeholders- investors, managers and analysts for making sound decisions.

The paper is relevant to the area of financial text analytics research because it provides a scalable and efficient framework in processing unstructured financial information. Future development can be based on real-time analytics and connection to financial decision support systems.

1. Introduction

1.1 Importance of Financial Reports in Business Decisions

The basis of corporate transparency and communication with stakeholders promote financial reporting. Annual reports, quarterly filings, management discussion and analysis (MD&A) applications and earnings call transcripts are all the most complete and authoritative data of the performance of a company, objectives, and second outlook. The documents will be the foremost foundation of information to a variety of parties, such as equity investors to review their positions in the portfolio, credit analysts to analyze the risks of default, regulatory authorities to ensure compliance with their regulation, and managers of the corporation to compare themselves with others in the respective industries.

The strategic worth of financial reports is far more than the arranged numerical tables that they consist of. Narrative disclosures such as management commentary relating to competitive positioning, descriptions of risk factors and future guidance are likely to hold the wealthiest and most useful intelligence, even though they are no longer accessible to automated intelligence-gathering systems. Those organizations that are able to quickly and properly interpret these stories are able to have a decisive information advantage in the financial markets.

1.2 Challenges of Analyzing Unstructured Text

Financial reports in their current form are daunting with respect to automated analysis in spite of their informational richness. An average SEC 10-K filing will be between 100 and 300 pages of formatted balance sheets and income statements which is filled with thick legal wading explanatory text. There are a number of features that make this text particularly hard to analyze:

- Volume and complexity: A single annual report may contain over 100,000 words across dozens of sections with highly specialized vocabulary
- Domain specificity: Financial language employs technical terminology ("amortization", "goodwill impairment", "covenant breach") whose meaning differs substantially from everyday usage
- Contextual ambiguity: Identical phrases carry opposite sentiment in different contexts - "above average" can be positive (performance) or negative (risk)
- Document heterogeneity: Reports vary significantly in structure, length, and writing style across companies, industries, and reporting periods
- Implicit information: Critical insights are often embedded in nuanced language, hedging constructions, and comparative references that require deep comprehension to interpret

1.3 Rise of AI and NLP in Finance

With the intersections of progress in Natural Language Processing (NLP), and the proliferation of more digitally-published financial materials, a novel challenge of automated financial text analysis has become an opportunity of an unprecedented scale. Use of AI systems in the financial industry is being actively implemented in applications such as textual sentiment analysis of trading information, compliance filings, forecasting earnings surprises, and credit risk estimation by textual disclosure as well as news sentiment analysis to give trading signals. Transformer revolution - launched by Vaswani et al. (2017) [13] and quickly followed by the likes of BERT, GPT and RoBERTa - has dramatically increased the ability of NLP systems to reason about complex, context unique language. Naturally trained networks, which learn representations based on billions of words of text, can transfer well to the domain of particular specialized task, with relatively little fine-tuning.

1.4 Limitations of Traditional Methods

Conventional techniques used to analyze financial text have been limited to rule-based and statistical techniques which are poorly adapted to the complexity of the financial language of today:

- Lexicon-based methods (e.g., Loughran-McDonald dictionary): Context-blind, static, cannot adapt to new terminology or negations
- Bag-of-words models with TF-IDF: Ignore word order and syntactic structure; suffer from high dimensionality and sparsity
- Support Vector Machines and Naive Bayes: Dependent on handcrafted features; poor generalization across document types
- Early neural networks (LSTM, RNN): Capture sequential structure but struggle with long-range dependencies and require large labeled datasets

These limitations result in systematic failures to capture the contextual, relational, and domain-specific dimensions of financial language that are most critical for insight extraction.

1.5 Problem Statement

Paper-based and other conventional automated approaches do not find rich and context-driven insights in the complex financial text. This is because of the absence of domain-specific knowledge, long document processing inability, and the inability to deal with subtle financial language leading to incomplete, inaccurate, and non-scalable financial intelligence. It is of paramount importance to have a smart, automated system that would be reliable in deriving actionable business information out of scale unstructured financial reporting

1.6 Objectives

The following are the main objectives to be used in this study:

1. Create a Bert-based NLP system adapted to financial report analysis
2. extraction- extract validation of major business insights on unstructured financial text, such as sentiment signals, risk factors and performance indicators
3. Compare the action of the proposed BERT-based system to the known baseline systems such as lexicon based models, conventional machine learning and previous deep-learning models
4. Discuss practical usefulness of extracted insights to the financial stakeholders such as investors, analysts and managers

2. Literature Review

This section is a review of the history of research of the early financial text analytics through the traditional methods of machine learning, deep learning methods, and finally transformer- based models where the proposed framework was built.

2.1 Financial Text Analytics

Computational finance of financial text has a long history that originates decades earlier. Initial studies acknowledged that predictive signatures were contained in textual information of financial reports and were not completely represented by numerical data. It is based on this foundational relationship amongst the tone of texts and stock market performance that the negative language present in Wall Street Journal columns is predictive of stock markets declines (Tetlock, 2007) [12]. Later endeavors were extended to the area of corporate disclosures. Loughran and McDonald (2011) [7] developed an important input by building the first lexicon of finance specific sentiment lexicon showing that negative dictionary words (using general purpose dictionary, e.g. Harvard General Inquirer) have neutral or technical senses in financial life. Their lexicon of words, such as positive, adverse, ambiguity, litigious and restricting words, was adopted as the standard of the financial NLP research almost ten years. Since then, NLP studies of finances have been developed to include various types of documents, such as quarterly earnings call transcripts (Qin and Yang, 2019) [10], analyst reports, or regulatory documents as

well as social media options. These have usage in return prediction and volatility forecasting to credit risk assessment and fraud detection.

2.2 Traditional Approaches

2.2.1 Lexicon-Based Approaches

Lexicon built approaches provide sentiment scores to documents by matching words in documents with predefined word lists and aggregating those scores at the sentence/document level. The most widely used lexicon that is cited in finance to gauge sentiment is Loughran- McDonald (LM) classical dictionary [7], the advanced dictionaries have since been created to process the language of the earnings calls, credit ratings statements, and central bank statements. Whereas lexicon based methods do create transparency, interpretability and are efficient in terms of computation, limitations are well documented. They are contextually blind in their essence, i.e. "outstanding" is a positive term whether they are speaking of an outstanding performance or outstanding liabilities. The aspects of negation (not profitable, no significant improvement), are either missing or restricted to mere negation windows. Premeditated vocabularies are not able to adjust to the changes of financial terms, changes in regulations language, and jargon usage.

2.2.2 Machine Learning Approaches (SVM, Naive Bayes)

Supervised machine learning models also overcame certain limitations of lexicon based methods by having classification limits learned over labelled training data and not using a set of predefined rules. TF-IDF features upon Support Vector Machines (SVMs) in the classification of financial text proved a formidable baseline in financial text classification tasks, and was proven competitive against sentiment classification and topic categorization benchmarks. Naive Bayes classifiers were based on the assumption of conditional independence and were computationally efficient and strong to small datasets, so they were used in the classification of financial news. Ensemble models such as the Random Forests and Gradient Boosting models have also been shown to achieve performance that is even better in structured sets of features that are based on both textual and non-textual cues. Nevertheless, these gains could not overcome the fact that the conventional machine learning models were limited in terms of requiring hand-crafted features, bag-of-words models that ignored the order of words in a document, and generalization of their applications across financial sectors and documents.

2.3 Deep Learning Approaches

2.3.1 Recurrent Neural Networks (RNN)

Sequential Processing NLP Recurrent Neural Networks added the concept of state representations to be hidden along position of tokens in models. On financial text, RNNs also learnt longitudinal dependencies, made them become more accurately captured than bag-of-words models. Nevertheless, vanilla RNNs had problem of vanishing gradient which inhibited their capability to detect long-range dependencies essential in the interpretation of complex financial sentences.

2.3.2 Long Short-Term Remembrance Networks (LSTM)

Long Short-Term Remembrance networks (Hochreiter & Schmidhuber, 1997) addressed the vanishing incline problem through fenced memory cells that selectively retain and reject information across long sequences. Bidirectional LSTMs (BiLSTMs) further extended this by processing systems in both forward and backward directions, providing richer contextual representations.

In earnings call transcript prediction, Qin and Yang (2019) [10] used LSTMs to predict stocks, and reported higher performance than lexicon-based baselines. As mechanisms of attention, models that are paired with LSTMs (BiLSTM-Attention) enabled them to dynamically assign different weights to contributions of the various positions of the sentence, enabling these models to enhance their financial classification performance. Nevertheless, LSTMs still had some weaknesses such as their sequential processing limitation, ineffectiveness in being able to parallelize training and still had problems with very long documents.

2.4 Transformer Models

2.4.1 BERT and Its Advantages

Devlin et al. (2019) [2] presented the concept of the publication of BERT (Bidirectional Encoder Representations from Transformers), which was a revolution in the field of NLP potential. The major innovations of Bert as compared to its predecessors are as follows:

- True bidirectionality: Unlike earlier models that treated text left-to-right or concatenated separate left and right passes, BERT's masked language modeling objective enables simultaneous attention to both directions
- Deep contextualization: Each word's representation is a function of all other words in the input, enabling disambiguation of polysemous terms and complex financial constructions
- Pre-training at scale: BERT is pre-trained on 3.3 billion words, acquiring broad linguistic knowledge that transfers to specialized tasks with nominal labelled data
- Fine-tuning efficiency: Task-specific adaptation requires only a lightweight classification head and a few epochs of training, dramatically reducing the labeled data requirement

Araci (2019) [1] demonstrated BERT's applicability to financial NLP through FinBERT, fine-tuned on the Financial PhraseBank sentiment dataset, achieving state-of-the-art accuracy of 86.2% — significantly outperforming LSTM and lexicon-based baselines. Yang et al. (2020) [14] also followed it up by further training on a 4.9 billion token financial corpus, raising the performance on various financial NLP benchmarks. Later developments In later transformer variants, such as RoBERTa (Liu et al., 2019) [6], DeBERTa (He et al., 2021) [3] and ALBERT (Lan et al., 2020) [5] added improvements to the architecture that further enhanced state-of-the-art results. The applicability of specialized pre-training on regulatory and audit text was shown by domain specific models such as SEC-BERT and AuditBERT.

2.5 Research Gap

Although it has to be admitted that significant progress has been achieved, three key gaps have been identified in the current literature: Lack of context-aware models: Even when lexicon-based models or general-purpose ML models are used, it is impossible to present domain-adapted financial text in the current literature; systematic methodologies of pre-trained financial corpus and vocabulary judgement remain unrealized Limited automation: Existing scientific studies have explored individual tasks (sentiment OR risk detection OR entity recognition) singly, but do not represent a comprehensive solution of extracting insights across multiple analysis.

3. Proposed Framework / Methodology

3.1 Overview of Framework

The proposed framework is an automated model of extracting business insights of unstructured financial reports. It consists of five consecutive steps, which are data collection, preprocessing, implementing the BERT model, extracting insights, and evaluating it. The pipeline will be modular too, meaning a simple stage is able to be updated or replaced, and can be scaled to massive document corpora.

The input phase takes the raw financial documents and ingests them in various sources. The preprocessing phase purifies, breaks down and partakes text into BERT. The independent Bert training phase considers an adaptive task-specific flattened training of domain pre-trained approaches. Four insight module models are used in the extraction stage. The output stage presents business intelligence in the form of structured and actionable information to the consumption by the stakeholders.

3.2 Data Collection

The framework ingests financial text from three primary source categories:

3.2.1 Annual Reports

The major source of data is annual reports (10-K filing in the US system). These are the Management Discussion and Investigation (MD&I) section, where the rich narrative is the performance of the business, its direction and future commentary, as well as the Item 1 (Business Description), the Item 1A (Risk Factors), and the Item 7A (Quantitative and Qualitative Disclosures About Market Risk). SEC EDGAR annual reports are in machine readable XBRL format allowing automated extraction of sections.

3.2.2 Financial Disclosures

Quarterly reports (10-Q) and earnings releases and call transcripts have increased frequency, which supplements annual report analysis. Transcripts of the earnings calls, including prepared statements by the management and question-and-answer portions with the analysts, are especially useful in terms of

literal temperature of management reading and proactive language that is not yet contained in the official filings.

3.2.3 Public Datasets

The framework is trained and evaluated using established public benchmark datasets, including the Financial PhraseBank [9] and FiQA 2018 [8] among others:

Table 1
Benchmark datasets used for training and evaluation

Dataset	Task	Size	Source
Financial PhraseBank	Sentiment (3-class)	4,845 sent.	Reuters financial news
FiQA 2018	Opinion QA	17,000 sent.	News & social media
SEC-FLS Corpus	FLS Detection	25,000 sent.	SEC EDGAR 10-K filings
FinNLP Risk Corpus	Risk Classification	10,000 para.	10-K Item 1A sections
EarningsCall QA	Extractive QA	8,000 pairs	S&P 500 earnings transcripts

3.3 Data Preprocessing

Raw financial text requires extensive preprocessing before it can be consumed by BERT. The preprocessing pipeline operates in the following sequence:

3.3.1 Text Cleaning

Financial documents contain substantial non-informative content that must be removed to improve signal quality. The cleaning stage applies the following operations:

- Boilerplate removal: Repeated legal disclaimers, safe harbor statements, table of contents entries, and digital signature blocks are detected and removed using regular expression patterns
- Table extraction: Structured numerical tables are identified and either excluded from text analysis pipelines or converted to descriptive text representations for hybrid analysis
- Special character normalization: Currency symbols ($, €, £), percentage signs, and financial shorthand ("bn", "mn", "bps") are standardized
- Whitespace and encoding normalization: Unicode normalization, smart quote standardization, and consistent line break handling

3.3.2 Tokenization

BERT employs WordPiece subword tokenization, which decomposes words into sub-word units from a learned 30,522-entry vocabulary. This approach handles out-of-vocabulary financial terms by decomposing them into recognizable subword components. For FinBERT, the vocabulary is augmented

with 500 high-frequency financial domain terms to reduce unnecessary fragmentation of important financial concepts.

Special tokens are inserted according to the BERT input format: [CLS] is prepended to mark the sequence start (its final representation is used for classification tasks), and [SEP] is inserted at sentence boundaries and document segment separations.

[CLS] token$_1$ token$_2$ ... token$_n$ [SEP]

3.3.3 Stopword Removal

Standard English stopwords are removed from inputs destined for keyword and key phrase extraction modules. Importantly, stopword removal is not applied for sentiment analysis and classification inputs, as function words and connectors ("not", "despite", "however") carry critical sentiment-modifying information in financial text. A finance-specific stopword list supplements the standard NLTK stopword set, excluding domain-common non-informative terms such as "fiscal", "quarter", "pursuant", and "thereof".

3.4 BERT Model Implementation

3.4.1 Pre-trained BERT

The backbone of the proposed framework is FinBERT, a BERT-Base model subjected to continued domain-adaptive pre-training on a 4.9 billion token financial corpus. BERT-Base comprises 12 Transformer encoder layers, 12 attention heads per layer, a hidden dimension of 768, and 110 million total parameters. The pre-training corpus for FinBERT includes Reuters financial news (1.8B tokens), SEC EDGAR 10-K/10-Q filings (2.5B tokens), and earnings call transcripts (0.6B tokens).

3.4.2 Fine-tuning Process

Task-specific fine-tuning adds a lightweight classification or token-labeling head on top of the frozen-then-gradually-unfrozen BERT encoder. The fine-tuning procedure follows a three-stage protocol:

1. Classifier warm-up (Epoch 1): Only the task-specific head is trained; BERT encoder weights are frozen. This prevents the large pre-trained representations from being disrupted by random initialization of the new head
2. Full fine-tuning (Epochs 2–4): All layers are unfrozen and trained jointly with a low learning rate (2e-5 to 5e-5) using linear warmup over 10% of steps followed by linear decay
3. Evaluation and selection: The checkpoint achieving best validation performance on the task-specific metric is retained

Table 2
Fine-tuning hyperparameters

Hyperparameter	Value
Optimizer	AdamW
Learning rate	2e-5 to 5e-5
LR schedule	Linear warmup + linear decay
Warmup proportion	10% of total training steps
Batch size	16 / 32
Training epochs	3 to 5
Weight decay	0.01
Dropout (classifier head)	0.1
Max sequence length	128 (sentences) / 512 (paragraphs)

3.4.3 Input Representation

BERT constructs its input representation by summing three distinct learned embeddings for each input token:

$E(t) = E_{token}(t) + E_{segment}(t) + E_{position}(t)$

- Token embeddings: Map each WordPiece token to a 768-dimensional vector via a learned embedding matrix
- Segment embeddings: Distinguish tokens belonging to sentence A (embedding A) from sentence B (embedding B) in two-segment inputs — used for sentence-pair tasks such as next sentence prediction and question answering
- Positional embeddings: Encode the absolute position of each token (0 to 511) as a learned 768-dimensional vector, providing the position-agnostic Transformer with sequence order information

The self-attention computation that processes these embeddings is defined as:

$Attention(Q, K, V) = softmax(Q K^{T} / \sqrt{d_k}) \cdot V$

where Q (query), K (key), and V (value) matrices are linear projections of the input embeddings, and $\sqrt{d_k}$ is a scaling factor. This mechanism allows each token to attend to all other tokens simultaneously, producing deeply contextualized representations.

3.5 Insight Extraction Modules

The framework implements four specialized insight extraction modules, each implemented as a distinct fine-tuned classification or extraction head on the shared FinBERT encoder:

3.5.1 Sentiment Analysis

The sentiment module will categorize the financial sentences and paragraphs into three groups, namely Positive, Neutral and Negative. A linear classification head is applied to the [CLS] token representation:

$P(y \mid x) = \text{softmax}(W_s \cdot h_{[CLS]} + b_s)$

where $W_s \in \mathbb{R}^{3\times768}$ and $b_s \in \mathbb{R}^3$. The module is optimized on Financial PhraseBank and a own-generated MD&A sentiment corpus. Averaged Process sentencing level predictions are contents of sentences are aggregated to document parts with confidence[4]altered averaging data resulting in document-wide, conversation-wide, and section-wide sentiment.

3.5.2 Risk Detection

Risk detection module recognizes and classifies risk-related words in financial disclosures, especially, 10-K Item 1A disclosures. It is posed as a multi-label classification scenario of seven risk categories:

- Market Risk: Exposure to price, interest rate, and currency fluctuations
- Credit Risk: Counterparty default and credit quality deterioration
- Liquidity Risk: Inability to meet financial obligations when due
- Operational Risk: Internal process failures, system outages, human error
- Regulatory/Legal Risk: Compliance failures, litigation, regulatory change
- Technology Risk: Cybersecurity, data privacy, system obsolescence
- Macroeconomic Risk: Recession, inflation, geopolitical instability

Independent sigmoid classifiers are applied per category to the [CLS] representation, trained with binary cross-entropy loss and class-frequency weighting to handle label imbalance.

3.5.3 Key Phrase Extraction

The key phrase decoding module is used to extract the most informative phrases in financial text to be summarized and indexed. It a combination of keyphrase scoring based on the importance of the tokens created by Bert with statistical keyphrase scoring:

- Token salience: Attention weights from the final BERT layer are aggregated to assign importance scores to individual tokens
- Phrase boundary detection: Adjacent high-salience tokens are merged into candidate phrases using a learned span classifier
- Keyphrase ranking: Candidate phrases are ranked by a combination of BERT salience score, TF-IDF weight, and position (MD&A opening paragraphs are prioritized)

3.5.4 Performance Indicators

The performance indicator module derives quantitative financial values stated in narrative text, an example of which is revenue growth, EBITDA margin, EPS, return on equity, and similar sales growth. This has been applied as a Named Entity Recognition (NER) task via BIO tagging:

$P(y_i \mid x) = \text{softmax}(W_{ner} \cdot h_i + b_{ner}) \quad \forall \text{ token } i$

Types of entities are: B-METRIC/I-METRIC metric name), B-MONEY/I-MONEY monetary value), B-PERCENT/I-PERCENT percentage figure), B-DATE/I-DATE reporting period and B-DIRECTION/I-DIRECTION increment/decrement. Obtained triplets of metric-values-period are organized in a table of performance indicators of individual documents.

3.6 Evaluation Metrics

The following measures are used to determine task performance and were chosen to have a complete picture of model capability in various analytical dimensions:

3.6.1 Accuracy

Measure of accuracy gives the ratio of correct classification of instances in all classes. Although intuitive, accuracy may be false when dealing with imbalanced data (e.g. when there are many Neutral sentences in sentiment classification). It is reported and accompanied with balanced measures of completeness.

$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$

3.6.2 Precision

The percentage of predicted positive cases, which are actually positive, is the measure of Precision - which is of crucial importance in finance as false positives (ex: a risk factor incorrectly marked as harmful) directly result into costs.

$\text{Precision} = TP / (TP + FP)$

3.6.3 Recall

Recall is the percentage of true positives correctly found - important where a false negative (including a risk factor that is missed) can be disastrous.

$\text{Recall} = TP / (TP + FN)$

3.6.4 F1-Score

The F1-score is a performance measure that gives a balanced performance of precision and recall, and it is applicable especially in the case of imbalanced financial NLP datasets. On minority classes, it is said that Macro-F1 (arithmetic mean of classes) is an indicator of their performance.

F1 = 2 × (Precision × Recall) / (Precision + Recall)

4. Results and Discussion

4.1 Model Performance Comparison

4.1.1 Sentiment Analysis Results

Table 3
Sentiment classification results on Financial PhraseBank dataset

Model	Accuracy	Precision	Recall	F1-Score
LM Dictionary Baseline	72.4%	69.3%	67.2%	68.1%
Naive Bayes + TF-IDF	75.8%	72.0%	70.9%	71.4%
SVM + TF-IDF	78.2%	75.6%	74.3%	74.9%
LSTM + GloVe	79.6%	77.1%	76.5%	76.8%
BERT-Base (fine-tuned)	88.3%	87.4%	86.5%	86.9%
FinBERT — Proposed	93.6%	93.0%	91.9%	92.4%

The proposed FinBERT model attains 93.6% accuracy and 92.4% F1-score which are a 21.2 percentage points higher than LM Dictionary baseline and 24.3 points higher than LM Dictionary baseline, respectively. The improvement in the best traditional baseline (SVM + TF-IDF) lies in both accuracy and F1-score at 15.4 points and 17.5 points, respectively. The comparison of BERT-Base (88.3%), with FinBERT (93.6%) shows that domain specific continued pre-training provides an extra 5.3% increase in accuracy, which proves adaptation strategy of financial corpus.

4.1.2 Risk Factor Classification Results

Table 4
Risk factor classification results on FinNLP Risk Corpus **4.1.3 Named Entity Recognition (Performance Indicators)**

Model	Accuracy	Precision	Recall	F1-Score
TF-IDF + SVM	71.2%	68.5%	66.1%	67.3%
BiLSTM + Attention	78.5%	75.8%	74.2%	75.0%
BERT-Base	83.7%	82.4%	81.0%	81.7%
FinBERT — Proposed	91.8%	91.2%	89.7%	90.4%

Table 5
Financial NER results (entity-level strict matching)

Model	Precision	Recall	F1-Score
CRF + Handcrafted Features	72.1%	68.4%	70.2%
BiLSTM-CRF	79.3%	76.8%	78.0%
BERT-Base + NER Head	85.6%	84.2%	84.9%
FinBERT + NER Head — Proposed	91.3%	90.1%	90.7%

4.2 Interpretation of Extracted Insights

In addition to the aggregate measures, the quality of extracted insights is determined by qualitative model output analysis on held-out samples of financial reports. Key observations include:

- Sentiment nuance: FinBERT correctly classifies sentences containing financial hedging language such as "while revenues improved modestly, margin compression remains a concern" as Mixed/Negative, while baseline models assign Positive sentiment based on the word "improved"
- Risk categorization: The multi-label risk classifier correctly assigns Technology Risk and Operational Risk labels to cybersecurity disclosures, a distinction that single-label baselines conflate
- Performance indicator extraction: The NER module successfully extracts metric-value-period triplets from complex constructions such as "year-over-year revenue growth of 12.4% for the fiscal year ended December 31, 2023" with 90.7% F1

5. Implications

5.1 Managerial Implications

5.1.1 Better Decision-Making

The suggested framework will convert raw financial text into structured, queryable intelligence that will be directly used to support management decision-making at various levels. Elevated executives are able to get automated corrosion of competitor MD&A trends which can foster quicker strategic reaction towards market uptakes. Without manual analyst analysis, investment committees can get risk-adjusted sentiment scores in all the portfolio company filings. In the case of board-level governance, risk factor disclosures are auto-monitored across peer companies in the industry, allowing the completeness of risk profiles to be benchmarked, and new risk themes to be identified, which are not covered by a company filing alone - a potent early warning system. Gradual decline in the tone of management is difficult to identify in a single reporting period; thus, the capability to monitor the sentiment trends of a variety of reporting periods allows for the identification of the gradual decline of the tone of the management, which could be the predecessor of earnings disappointments.

5.1.2 Risk Identification

The multi-label risk detecting module is especially useful in the enterprise risk management functions. The framework makes use of automatic classification and quantification of risk disclosures among large portfolios of filers and allows:

- Portfolio-level risk screening: Identify all companies in a portfolio with elevated Liquidity Risk or Regulatory Risk disclosures without manual reading
- Risk trend analysis: Track the frequency and severity language of specific risk categories across reporting periods to detect deteriorating risk profiles
- Peer comparison: Benchmark a company's risk disclosure completeness against industry peers, identifying potential undisclosed risks
- Regulatory monitoring: Flag new regulatory risk language in real-time as 10-K filings are published on SEC EDGAR

5.2 Practical Implications

5.2.1 Automation in Financial Analytics

The framework provides high productivity in financial analytics processes, which are now reliant on document-by-document review. One average financial analyst takes 3–5 hours to look through a single 10-K filing to get sentiments and risk indicators. This is brought down to minutes in the proposed framework during initial screening where the analogous analyst is able to add value to the edge cases and model output validation and more intricate synthesis that actually needs human judgment. On scale, the framework allows covering the entire universe of public company filings - more than 10,000 10-K filings filings each year in the US alone - as opposed to the small subset that manual processes can practically handle. This democratizes the access to deep financial intelligence which results in less information asymmetry between the large institutional stakeholders with the dedicated research team and smaller market participants.

5.2.2 Use in Fintech Systems

The modular architecture of the proposed framework makes it well-suited for integration into fintech products and financial data platforms:

- Credit scoring platforms: Risk factor and sentiment signals from borrower financial disclosures can augment traditional quantitative credit models
- ESG analytics: Framework adaptation for ESG-related disclosure extraction supports sustainable finance analytics and regulatory reporting (e.g., TCFD climate risk disclosures)
- Regulatory technology (RegTech): Automated compliance monitoring of regulatory filing language supports financial institution compliance functions

- Investor relations platforms: Automated sentiment benchmarking of earnings communications helps IR teams calibrate messaging tone against market expectations
- Robo-advisory systems: Sentiment and risk signals from financial reports can inform portfolio rebalancing signals in automated investment platforms

6. Conclusion

The present study suggested and tested a multifaceted BERT-based architecture to extraction of the actionable business insights out of unstructured financial reports. The framework seals an essential knowledge gap in financial analytics: the inadequacy of traditional NLP and machine learning methods to identify in-depth context-based conclusions within the intricate textual data of annual reports, MD&A sections, risk reports and earnings communications.

6.1 Summary of Findings

Results of experimental outcomes of three main tasks indicate significant and consistent increases in experimental programs compared to baselines:

- Sentiment analysis: FinBERT achieves 93.6% accuracy on Financial PhraseBank, outperforming the best traditional baseline (SVM + TF-IDF at 78.2%) by 15.4 percentage points
- Risk factor classification: FinBERT achieves 91.8% accuracy on multi-label risk classification, outperforming BiLSTM-based approaches by 13.3 points
- Financial NER: FinBERT achieves 90.7% entity-level F1 on performance indicator extraction, outperforming BiLSTM-CRF by 12.7 points

6.2 Importance of BERT in Financial Analytics

These outcomes are in agreement with the theory that the bidirectional contextual pre-training of BERT is a genuine paradigm shift of financial text analytics. It is the only model that can deal with the vagaries of financial language due to its capacity to preserve the context-dependent semantics of a word, its capacity to operate with domain-specific vocabulary and its capacity to extrapolate limited labelled data. Domain-specific continued pre-training (FinBERT) shows similar extra performance when compared to the general-domain BERT, which supports the importance of matching pre-training data distribution with domain properties.

6.3 Key Contributions

The research contributes to the financial NLP literature in the following ways:

An end-to-end BERT-based system, which combines sentiment analysis, risk detection, key phrase extraction, and performance indicator extraction into a single coherent system, the first such system of its kind to be described in literature

A systematic comparison of B Bert and FinBert with results of lexicon-based, traditional ml, and deep learning baselines in a variety of financial NLP benchmarks, a comprehensive performance landscape

Applied explanation of the business value of the framework using the case studies of financial reports, filling the gap between research in NLP and practice in the field of financial analytics

A modular implementation architecture that can be extended to new financial document types, languages, and tasks of all types of analysis.

7. Future Work

Although the suggested framework shows high outcomes in the target tasks, multiple significant directions are still left by the future research and development:

7.1 Real-Time Analysis

The existing model is a batch model and the filings are processed after they are published. The further work will involve a creation of streaming ingestion pipeline, which follows SEC EDGAR in real-time to automatically set on it an analysis which will be performed within minutes after a new publication of a filing. This necessitates the latency of the inference time - now 45 seconds per 10 A100 GPUs - to be reduced to less than 10 seconds by model quantization (INT8), knowledge distillation (DistilFinBERT) [11], and batching. Real-time analysis would facilitate the trading signals due to time sensitivity like earnings releases and 8-K material material event disclosures.

7.2 Integration with Dashboards

This structured output of the framework will be incorporated together with the gadgets of financial analytics to the interactive add-on of business intelligence interfaces to the non-technical stakeholders. Planned features include:

- Time-series visualization of sentiment trends across reporting periods for individual companies and industry sectors
- Interactive risk heat maps showing risk category intensity across portfolio companies
- Automated report summaries combining key phrase extraction with abstractive summarization using generative LLMs
- Comparative analysis tools enabling peer benchmarking of risk disclosure completeness and sentiment positioning
- Alert systems that notify analysts when a company's filing sentiment or risk profile deviates significantly from historical patterns or peer norms

7.3 Use of FinBERT and Advanced Models

Subsequent model development will discuss a number of architectural improvements on the current implementation of FinBERT:

- FinBERT-Large: Scaling to the 24-layer, 340M parameter BERT-Large configuration for tasks where additional model capacity is expected to yield gains
- DeBERTa-Financial: Adapting DeBERTa's disentangled attention mechanism — which separately represents content and position — for financial text, where precise positional relationships between financial entities are semantically significant
- LLM integration: Combining the extractive precision of fine-tuned BERT models with the generative capabilities of large language models (GPT-4, Claude) for abstractive financial report summarization
- Cross-lingual extension: Training multilingual FinBERT variants on non-English financial corpora (EU IFRS filings, Japanese securities reports, Chinese annual reports) to support global financial analytics
- Temporal modeling: Incorporating longitudinal document sequences to model how company language evolves over time, enabling early detection of deteriorating financial narratives

Declarations

Author Contribution

Dr. Prasant Kumar RoutAssistant Professor (IT), Sri Sri University, Cuttack, OdishaEmail: prasant.rt@gmail.comORCID: 0009-0002-4938-6170Dr. Prasant Kumar Rout contributed to the conceptualization of the study, design of the research framework, methodology development, data preprocessing, model implementation, and manuscript drafting. He played a key role in integrating the BERT-based architecture and performing the experimental analysis.Ms. Binita NandaAssistant Professor (Finance), Sri Sri University, Cuttack, OdishaEmail: binita.nanda18@gmail.comORCID: 0000-0002-3493-4114Ms. Binita Nanda contributed to domain-specific insights, financial data interpretation, and validation of extracted business insights. She also assisted in literature review, formulation of financial context, and refinement of the results from a financial perspective.Dr. Sunil Kumar DhalProfessor (IT), Sri Sri University, Cuttack, OdishaEmail: sunildhal@srisriunivesity.edu.inORCID: 0000-0003-2324-7092Dr. Sunil Kumar Dhal provided overall supervision, guidance on research design, and critical review of the manuscript. He contributed to methodological validation, result interpretation, and final editing of the paper.All authors reviewed and approved the final version of the manuscript and agree to be accountable for all aspects of the work.

References

1. Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063.

2. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, 4171–4186.
3. He, P., Liu, X., Gao, J., & Chen, W. (2021). DeBERTa: Decoding-enhanced BERT with disentangled attention. International Conference on Learning Representations (ICLR 2021).
4. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.
5. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A lite BERT for self-supervised learning of language representations. ICLR 2020.
6. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.
7. Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. Journal of Finance, 66(1), 35–65.
8. Maia, M., Handschuh, S., Freitas, A., Davis, B., McDermott, R., Zarrouk, M., & Balahur, A. (2018). WWW'18 open challenge: Financial opinion mining and question answering. Companion Proceedings of WWW 2018, 1941–1942.
9. Malo, P., Sinha, A., Korhonen, P., Wallenius, J., & Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. Journal of the Association for Information Science and Technology, 65(4), 782–796.
10. Qin, Y., & Yang, Y. (2019). What you say and how you say it matters: Predicting stock volatility using verbal and vocal cues. Proceedings of ACL 2019, 390–401.
11. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.
12. Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. Journal of Finance, 62(3), 1139–1168.
13. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.
14. Yang, Y., Uy, M. C. S., & Huang, A. (2020). FinBERT: A pretrained language model for financial communications. arXiv preprint arXiv:2006.08097.

Figures



Figure 1

High-level pipeline of the proposed BERT-based financial insight extraction framework