FedMat: A Federated Multimodal Materiality Framework for Trustworthy Financial Document Analysis

Filip Habdas[1], Liyuan Liu[2]
[1]Department of Mathematics, Saint Joseph's University
[2]Department of Decision and System Sciences, Saint Joseph's University
5600 City Ave, Philadelphia, PA 19131, USA
fhabdas@sju.edu, lliu@sju.edu

Abstract

Financial document analysis has become increasingly essential because it supports business decision-making and investment decisions. However, financial documents have unique characteristics, where the style, tone, and financial indicators can substantially influence the results of the analysis. At the same time, a major challenge in financial document analysis is data security. To address this issue, we develop a new Federated Multimodal Materiality framework, named FedMat, which incorporates not only textual data but also materiality-related signals, including percentage change, monetary amount, qualitative category, and contextual and linguistic features. In addition, FedMat operates in a federated learning environment. Based on our experiments, combining text with materiality scores consistently outperforms text-only approaches for financial text classification, particularly for nonlinear models. In general, FedMat achieves the strongest performance, with AUCs of 90.2%-91.3%, compared to 83.6%-90.3% for text-only methods. The greatest gain is observed for XGBoost, improving by 7.7%. In federated learning settings, FedMat maintains strong utility under mild adversarial conditions, while simple robust aggregation provides an effective and practical defense that improves resilience against stronger poisoning attacks.

Introduction

Financial document analysis is important in modern business and investment decision-making because it helps firms and stakeholders understand market conditions and evaluate company performance. Financial disclosures, such as annual reports, earnings announcements, and news releases, provide valuable qualitative information that can shape how investors interpret a company's performance. However, traditional financial indicators alone cannot fully capture the subtle signals embedded in managerial communication, such as tone, emphasis, or the use of optimistic or pessimistic language. For this reason, text-based analysis has become increasingly useful for identifying emotional and linguistic signals and for understanding how financial information is communicated and perceived. For example, Loughran and McDonald (Loughran and McDonald 2011) report that a more negative tone in 10-K filings is significantly associated with abnormal stock returns of approximately 1.5% in the three days following disclosure.

However, text-based analysis also has limitations, and relying solely on linguistic cues can lead to incomplete or biased conclusions. Unlike many other text sources, financial documents possess unique characteristics. Sentiment detection in financial text requires materiality variables in addition to language features, because linguistic indicators alone do not convey the economic significance of disclosed information. The economic impact of a disclosure depends on both context and magnitude, as financial sentiment inherently involves scale and relevance. For example, a $5 million loss may have little material effect on a multinational corporation, but could trigger a major liquidity crisis for a small-cap firm. Even models using advanced contextual embeddings cannot infer the economic magnitude of language alone. Evaluating the impact of disclosures on investor decisions and market value therefore depends on materiality variables such as firm size, revenue, leverage, volatility, segment importance, and year-over-year performance changes. Previous studies (van der Lugt, Bakker, and Mans-Kemp 2025; Kinney, Burgstahler, and Martin 2002; Liao 2025) show that market reactions arise from the interaction between textual sentiment and financial materiality rather than sentiment polarity alone. Consequently, financial sentiment models achieve stronger explanatory and predictive performance when materiality features are incorporated, as these features align with the decision criteria used by investors, analysts, and regulators.

In addition to the unique characteristics of financial documents, another major challenge in the financial sector is the risk of data breaches. As Artificial Intelligence (AI) becomes more widely adopted and models require increasingly large datasets for training, concerns about data security and privacy become even more critical. For example, in July 2019, Capital One revealed that a hacker gained unauthorized access to its systems, compromising the personal information of more than 100 million individuals (Khan et al. 2022). Similarly, in 2017, Equifax suffered a breach that exposed the personal data of 147 million U.S. consumers, including names, Social Security numbers, and dates of birth (Dias, Borbinha, and Mendonça 2022). Data breaches can be extremely damaging. The Equifax breach alone resulted in regulatory settlement costs of US $575–700 million, while

the Capital One breach incurred an additional US $100–150 million in remediation expenses in its first year (Federal Trade Commission (FTC) 2024; Capital One Financial Corporation 2019). Beyond monetary losses, data breaches also cause substantial non-financial harm, including reputational damage, loss of customer trust, and long-term operational risks. Therefore, when applying AI to financial document analysis, including tasks such as sentiment analysis, data security becomes a critical consideration. Federated learning provides a decentralized framework in which models are trained locally on distributed financial datasets, while only model parameters, rather than raw data, are transmitted to a central aggregator (Liu et al. 2025, 2023). This approach reduces the risk of exposing sensitive information while preserving the analytical performance of AI models.

To address the two major limitations in financial document sentiment analysis, namely the inability to rely solely on textual information and the growing concern over data security, we propose a framework named FedMat, a federated multimodal materiality framework designed to support trustworthy financial sentiment analysis. In particular, the contributions of this study are as follows.

- To address the limits of text-only document analysis in financial documents, FedMat introduces a materiality-aware framework that integrates quantitative financial indicators with contextual language features. We construct a data driven materiality score that merges percentage changes, monetary magnitudes, key financial categories, and signals. A classification layer learns an integrated MatScore from these features, enabling FedMat to identify economically meaningful disclosures and improve the reliability of sentiment prediction.
- To address the critical limitation of data security in financial document analysis, FedMat incorporates a federated learning environment that enables collaborative model training across multimodal financial datasets while ensuring that raw text and numerical information remain local.
- We conducted experiments using six model configurations, including embedding-based classifiers (sentence embeddings with Logistic Regression, calibrated Linear SVM, XGBoost, and MLP) as well as transformer fine-tuning models (DeBERTa, and GPT-2). For each model, we evaluated four settings: text-only, FedMat, federated text-only, and federated FedMat. Our results demonstrate the feasibility of FedMat, its consistent performance improvements over traditional text-only baselines, and its effectiveness when deployed in a federated learning environment.

The remainder of the paper is organized as follows: Section 2 reviews related work on materiality, multimodal financial document analysis, and federated trustworthy learning. Section 3 introduces the FedMat framework and materiality scoring design, and Section 4 reports the experimental settings and results under centralized and federated deployments.

Related Works

Materiality and Information Value in Financial Disclosure

The study by Anderson Rafael Costa Sousa (Sousa, Pacheco, and Rover 2024) examines how the level of disclosure and the materiality of intangible assets affect their relevance in value (i.e., their impact on market value) in publicly traded Brazilian companies. The results show that more material disclosure reduces information asymmetry and enables investors to make more accurate pricing decisions. More broadly, higher information quality and transparency improve predictive power in financial markets, highlighting the central role of material information in investor decision-making. Material sustainability disclosures (Grewal, Hauptmann, and Serafeim 2021) similarly make stock prices more informative, as shown by Grewal, Hauptmann, and Serafeim. When disclosures are non-material, they have little effect on stock prices. Their findings support the development of ESG reporting frameworks that emphasize material sustainability issues in disclosure requirements. These studies collectively reinforce the idea that materiality is tied to information value rather than moral importance; information is considered material only if it improves investor decision-making (Mosca and Picciau 2020). Similarly, Langevoort (Langevoort 1998) explains that financial disclosure prioritizes economically relevant information because its primary purpose under securities law is investor protection. Information must be disclosed only when it has a clear financial impact. From another perspective, stakeholder and legitimacy theories (Pizzi, Principale, and De Nuccio 2023) suggest that materiality assessments often lack detailed stakeholder prioritization and may function more as legitimacy tools than as mechanisms for identifying useful information for decisions. This suggests that sustainability materiality can sometimes be disconnected from actual decision usefulness, reinforcing the need for analytical methods that explicitly model economically meaningful signals.

Taken together, this literature indicates that materiality is central to how financial information affects markets, yet prior studies largely rely on centralized statistical analyzes or qualitative evaluation. This motivates our approach: instead of treating materiality as a reporting concept, we operationalize it as a predictive signal in financial sentiment modeling. Furthermore, because material financial data are often sensitive and distributed between firms, federated learning provides a practical framework for integrating this information while preserving privacy. By combining materiality-sensitive features with distributed learning that preserves privacy, our framework links disclosure theory with modern machine learning to improve both predictive performance and secure data utilization in financial analysis.

AI and Multimodal Approaches in Financial Document Analysis

Cho et al. (Cho et al. 2023) propose a framework for analyzing unstructured financial documents using robotic process automation (RPA) and multimodal AI models. In their

pipeline, RPA collects documents, forwards them to an AI model for interpretation, and then formats and uploads the extracted information. The model used in their study, LayoutXLM, is a multilingual multimodal transformer designed to process both textual and visual elements in structured documents. This work illustrates how automation and multimodal modeling can improve the extraction of financial information from complex document formats. Similarly, Pappula and Rusum (Pappula and Rusum 2023) develop a multimodal document understanding framework that integrates optical character recognition, transformer-based natural language processing, and computer-vision-based layout analysis to jointly process textual, visual, and structural information. Their results highlight that AI can serve as a filtering mechanism that transforms raw disclosed data into structured and decision-useful financial information.

Building on this line of work on automated multimodal document processing, our study extends the pipeline by introducing a federated learning framework to enable large-scale, privacy-preserving model training. While prior approaches focus primarily on document extraction and structuring, federated learning allows multiple data owners to collaboratively train multimodal models without centralizing sensitive raw data. This design not only improves scalability and supports the training of larger models but also enhances the reliability and stability of model estimation under realistic data governance and security constraints.

Federated and Trustworthy Learning in Financial AI

Recent studies have explored federated learning in financial risk analysis. For example, Aljunaid et al. (Aljunaid et al. 2025) apply federated learning to train fraud detection models across multiple banks without sharing raw transaction data, and incorporate explainable AI techniques to improve model transparency. Tang et al. (Tang and Liang 2024) propose a federated graph-learning-based credit card fraud detection framework that integrates federated learning with graph neural networks to support privacy-preserving multi-institution collaboration. Similarly, Alhasawi et al. (Alhasawi, Almtrf, and Asad 2025) introduce FedFraud, a federated fraud detection framework that combines trust-aware client aggregation with asynchronous communication to improve robustness under non-IID data conditions. Their results show a stronger detection performance while reducing privacy leakage. In addition, Aljunaid et al. (Aljunaid et al. 2025) develop an Explainable Federated Learning (XFL) framework that integrates federated learning with explainability methods such as SHAP and LIME, allowing privacy-preserving and interpretable fraud classification with reduced false positives. Mosaiyebzadeh et al. (Mosaiyebzadeh et al. 2024) used federated learning in IoHT environments combining with differential privacy to build secure intrusion detection systems for healthcare devices, demonstrating that collaborative training can improve attack detection while protecting sensitive patient data.

However, existing research primarily focuses on federated learning for traditional financial risk tasks such as fraud detection and credit scoring, or on general trust and robustness mechanisms in federated systems. Federated multimodal financial sentiment analysis that explicitly incorporates economic materiality signals remains largely unexplored. This gap motivates our framework, which integrates multimodal financial text analysis with materiality-aware features in a federated environment that preserves privacy.

Methodology

Overview of the FedMat

FedMat consists of three layers: the Materiality Scores Layer, the Federated Learning Layer, and the Business Decision-Making Layer. As illustrated in Figure 1, the Materiality Scores Layer collects non-textual information from financial institutions and computes the final materiality score. After the materiality scores are generated, the Federated Learning Layer collaboratively trains a global model across decentralized clients without sharing raw data. Specifically, each client updates its local model using the extracted multimodal representations and the corresponding materiality scores, and only the model parameters are uploaded to the server for secure aggregation. The aggregated global model is then broadcast to all clients for the next training round, allowing privacy-preserving learning while improving model generalization in financial document analysis. The resulting analytical outputs are finally passed to the Business Decision-Making Layer to support downstream decision-making.

Materiality Layer: Definition and Design

Our FedMat framework proposes a sentence-level materiality scoring mechanism that integrates both quantitative and qualitative indicators, motivated by SEC SAB 99 (Fang and Jacobs 2000) and IFRS materiality principles. For each disclosure sentence s, an engineered feature representation is constructed to capture (i) relative percentage-change signals, (ii) scale and monetary magnitudes, (iii) categories of financially material events, and (iv) contextual linguistic characteristics. These features are used to train a supervised logistic regression model that produces a probabilistic materiality score $\mathrm{MatScore}(s)$. The materiality-related information extracted from each sentence includes the following components:

(1) Percentage-Change Component $P(s)$: Let $\mathcal{P}(s) \subseteq \mathbb{R}_{\geq 0}$ denote the set of percentage-change magnitudes extracted from sentence s. To ensure a well-defined computation when no percentage is present, we define

$$p_{\max}(s) = \max\left(\{0\} \cup \mathcal{P}(s)\right).$$

Consistent with the emphasis on relative deviations in SAB 99, the percentage-change component is defined as

$$P(s) = \min\left(1,\ \frac{p_{\max}(s)}{\tau_P}\right),$$

where $\tau_P > 0$ is a scaling and saturation threshold that maps percentage deviations to the unit interval. In this study, τ_P is set to 20, such that percentage changes of 20% or greater are saturated at $P(s) = 1$, while smaller deviations are linearly scaled.

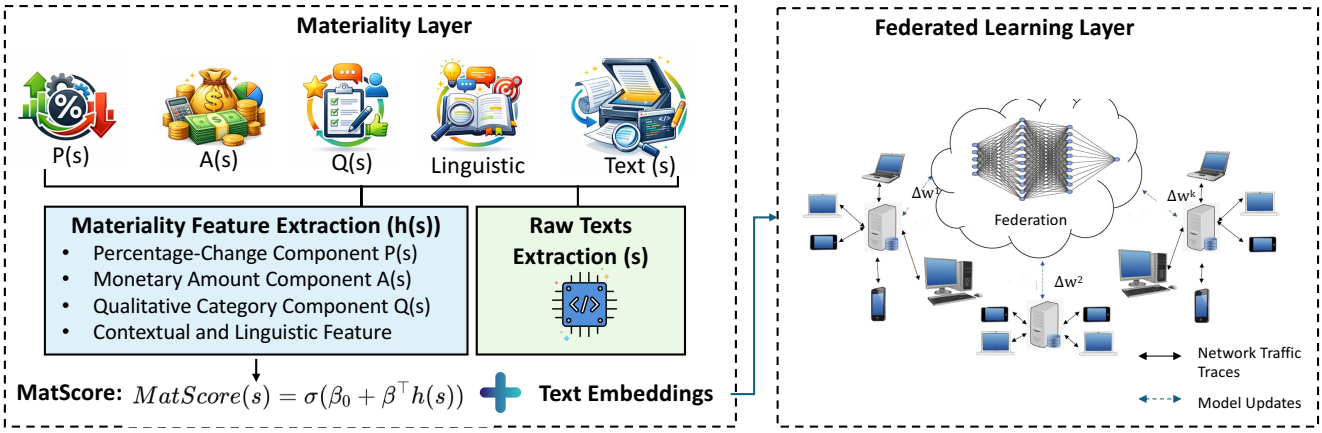

Figure 1: Framework of FedMat

(2) Monetary Amount Component $A(s)$: For each sentence s, all monetary amounts are normalized to millions of USD. For instance, \$2 billion is mapped to 2000, and \$500k is mapped to 0.5. If multiple monetary amounts occur in the same sentence, the maximum value remains because in financial documents, the largest magnitude typically contributes most to materiality. Formally, let $\mathcal{A}(s) \subseteq \mathbb{R}_{\geq 0}$ denote the set of normalized amounts extracted from s, and define

$$a_{\max}(s) = \max\left(\{0\} \cup \mathcal{A}(s)\right).$$

To model diminishing sensitivity to large monetary magnitudes, the monetary component is defined as

$$A(s) = \min\left(1,\ \log_{10}\left(1 + \frac{a_{\max}(s)}{\tau_A}\right)\right),$$

where $\tau_A > 0$ is a scaling parameter. In this study, τ_A is set to 10, corresponding to a reference scale of \$ 10M under normalization in millions of dollars. Transformation $\log_{10}(1+x)$ compresses the dynamic range of large monetary values and reduces the influence of extreme outliers, while the saturation operator $\min(1, \cdot)$ ensures that $A(s) \in [0, 1]$.

(3) Qualitative Category Component $Q(s)$: A set of financially material event categories is detected from each sentence s, including *profit, loss, revenue, M&A, risk, debt*, and *employment-related* events. Let $\mathbf{z}(s) \in \{0,1\}^7$ denote the category indicator vector, where each entry corresponds to one category. The qualitative category signal is constructed as a weighted aggregation motivated by previous accounting and finance studies on value-relevant disclosures and market reactions (Kothari 2001; Jegadeesh and Livnat 2006; Andrade, Mitchell, and Stafford 2001; Campbell et al. 2014; Frank and Goyal 2003; Edmans et al. 2024). Specifically, let

$$\mathbf{w} = \left[0.25,\ 0.25,\ 0.20,\ 0.20,\ 0.20,\ 0.10,\ 0.05\right]^\top,$$

where the entries correspond to *profit*, *loss*, *revenue*, *deal*, *risk*, *debt*, and *employment*, respectively. The qualitative category component is then defined as

$$Q(s) = \min\left(1,\ \mathbf{w}^\top \mathbf{z}(s)\right).$$

This component serves as an interpretable qualitative signal that captures the presence of financially significant categories. It is not used as a prediction target; instead, the final materiality score is learned from the data by the supervised model using the full feature representation. Although category weights and thresholds are informed by prior literature and data characteristics, preliminary perturbation tests suggest that FedMat's performance is stable and a more comprehensive sensitivity and ablation analysis will be explored in future work.

(4) Contextual and Linguistic Features: In addition to the numerical and category-based components, we extract a set of contextual and linguistic indicators that capture the disclosure style and informational content of each sentence s. These features include forward-looking language, backward-looking descriptions, hedging and uncertainty indicators, negation markers, monetary size-threshold flags, finance-domain sentiment categories (Loughran–McDonald), finance-specific keyword intensity, and a numerical change-direction indicator. Let $\mathbf{c}(s) \in \mathbb{R}^d$ denote the contextual feature vector, where the following components are included. *Forward-looking language* ($c_{\text{fwd}}(s)$) indicates whether the sentence contains predictive or planning-oriented expressions such as *"expect"*, *"will"*, *"anticipate"*, and *"forecast"*. *Backward-looking descriptions* ($c_{\text{bwd}}(s)$) indicate whether the sentence reports realized outcomes or historical performance (e.g., *"in the prior year"*, *"during 2024"*). *Hedging* ($c_{\text{hedge}}(s)$) flags probabilistic or softened language such as *"may"*, *"could"*, *"might"*, and *"potentially"*, which is frequently used in SEC filings when discussing uncertain impacts. *Uncertainty cues* ($c_{\text{uncert}}(s)$) capture explicit indicators of ambiguity or unpredictability (e.g., *"volatility"*, *"unpredictable"*). *Negation* ($c_{\text{neg}}(s)$) flags negation markers such as *"not"*, *"no"*, and *"never"*. *Size-threshold flags* ($c_{\geq 10m}(s)$ and $c_{\geq 100m}(s)$) encode whether the sentence contains monetary magnitudes exceeding \$10 million or \$100 million, respectively, after unit normalization. Loughran–McDonald *sentiment indicators* ($c_{\text{LM-pos}}(s)$ and $c_{\text{LM-neg}}(s)$) capture the presence of pos-

itive and negative sentiment words in the finance domain. *Finance keyword intensity* ($c_{\text{fin-int}}(s)$) quantifies the density of high-salience finance-related keywords such as *"material"*, *"significant"*, *"restatement"*, *"impairment"*, and *"litigation"*. Finally, the *change-direction code* ($c_{\text{dir}}(s)$) is an integer-valued indicator capturing the direction of numerical changes mentioned in the sentence, encoded as -1, 0, 1, and 2 for downward, flat, upward, and mixed changes, respectively.

Each sentence is represented by the concatenated feature vector

$$\mathbf{h}(s) = \big[P(s),\, A(s),\, Q(s),\, \mathbf{c}(s)^\top\big]^\top.$$

In the implementation used for this study, we set $d = 11$.

(5) Learning a Data-Driven Materiality Score: Given binary labels $y(s) \in \{0, 1\}$ indicating whether sentence s contains material disclosure information, we fit a supervised logistic regression model on the engineered feature vector $\mathbf{h}(s)$. The sentence-level materiality probability is defined as

$$\text{MatScore}(s) = \sigma\Big(\beta_0 + \boldsymbol{\beta}^\top \mathbf{h}(s)\Big), \qquad \sigma(x) = \frac{1}{1 + e^{-x}}.$$

For interpretability, the final materiality score is reported on a scale of 0 to 100:

$$\widehat{M}(s) = 100 \cdot \text{MatScore}(s).$$

Model performance is evaluated using 5-fold stratified cross-validation, reporting precision, recall, F1-score, AUC, and accuracy. Statistical uncertainty is quantified via bootstrap 95% confidence intervals for AUC and Brier score using 1000 resamples. The algorithm 1 shows the process for constructing the MatScore. It is a normalized characteristic vector for each sentence by combining percentage changes, monetary magnitudes, categorical indicators, and contextual cues, and then fits a logistic regression model to produce a final sentence-level materiality score.

Federated Learning in FedMat

Federated learning enables collaborative model training across multiple data holders without sharing raw data, making it well suited for privacy-sensitive financial analysis (Liu and Han 2024). Instead of centralizing data, each client trains a local model on its proprietary data set and shares only model updates with the coordinating server, which aggregates them to form a global model. FedMat naturally supports this paradigm because its feature extraction and materiality scoring are performed locally, ensuring consistent representations across institutions while preserving data confidentiality. Because financial materiality varies across industries, firm sizes, and reporting contexts, aggregating learning signals from multiple clients allows the model to capture broader patterns in economically meaningful disclosures and improves generalization across heterogeneous environments. In our experiments, the dataset is partitioned into simulated clients $K = 5$, each training a local model for several epochs (15 epochs for deep models), and the server aggregates client outputs using sample-size–weighted averaging to approximate cross-institution collaboration while preserving data privacy.

Algorithm 1: Sentence-Level Materiality Score

Require: Labeled sentences $\{(s_i, y_i)\}_{i=1}^N$, $y_i \in \{0, 1\}$
Require: Extractors: $\text{Amounts}(\cdot)$, $\text{Assets}(\cdot)$, $\text{Cat}(\cdot)$, $\text{Ctx}(\cdot)$
Require: Category weights $\mathbf{w} \in \mathbb{R}^7_{\geq 0}$ with $\|\mathbf{w}\|_1 \leq 1$
Require: Scaling constants $\tau_P > 0$, $\tau_A > 0$
Ensure: Materiality scores $\{\widehat{M}(s_i)\}_{i=1}^N$ and parameters $(\beta_0, \boldsymbol{\beta})$

Define $\text{clip}(x) = \min\{1, \max\{0, x\}\}$
for $i = 1$ to N **do**
  $\mathcal{V}_P \leftarrow \text{Amounts}(s_i) \subseteq \mathbb{R}_{\geq 0}$
  $p_{\max} \leftarrow \max(\{0\} \cup \mathcal{V}_P)$
  $P(s_i) \leftarrow \text{clip}\Big(\frac{p_{\max}}{\tau_P}\Big)$
  $\mathcal{V}_A \leftarrow \text{Assets}(s_i) \subseteq \mathbb{R}_{\geq 0}$
  $a_{\max} \leftarrow \max(\{0\} \cup \mathcal{V}_A)$
  $A(s_i) \leftarrow \text{clip}\Big(\log_{10}\Big(1 + \frac{a_{\max}}{\tau_A}\Big)\Big)$
  $\mathbf{z}(s_i) \leftarrow \text{Cat}(s_i)$ ▷ $\mathbf{z}(s_i) \in \{0, 1\}^7$
  $Q(s_i) \leftarrow \text{clip}\big(\mathbf{w}^\top \mathbf{z}(s_i)\big)$
  $\mathbf{c}(s_i) \leftarrow \text{Ctx}(s_i)$ ▷ $\mathbf{c}(s_i) \in \mathbb{R}^d$
  $\mathbf{h}(s_i) \leftarrow [P(s_i), A(s_i), Q(s_i), \mathbf{c}(s_i)^\top]^\top \in \mathbb{R}^{3+d}$
end for
Fit a (regularized) logistic regression:

$$\text{MatScore}(s) = \sigma(\beta_0 + \boldsymbol{\beta}^\top \mathbf{h}(s)), \quad \sigma(x) = \frac{1}{1 + e^{-x}}$$

for $i \leftarrow 1$ to N **do**
  $\widehat{M}(s_i) \leftarrow 100 \cdot \text{MatScore}(s_i)$
end for
return $(\beta_0, \boldsymbol{\beta}, \{\widehat{M}(s_i)\}_{i=1}^N)$

Experimental Results and Discussion

Dataset Description and Experimental Settings

We used a public FinancialPhraseBank dataset to run the experiments downloaded from Kaggle (Malo et al. 2014). This dataset contains the sentiments for financial news headlines from the perspective of a retail investor. In this data set, each record includes two fields: "News Headline", which provides a brief financial news statement, and "Sentiment", which indicates the investor oriented emotional polarity of the headline as positive, neutral, or negative.

Materiality Layer: For the experimental setup, we evaluated seven combinations of embedding and classification models for sentiment classification in the original data set. For each combination of models, we conducted four experiments: (1) text-only, (2) FedMat, (3) federated text-only, and (4) federated FedMat. For word embedding, we extract dense sentence embeddings using SBERT (all-MiniLM-L6-v2) when the model weights are locally available. If SBERT is unavailable, we apply a fully offline fallback embedding pipeline using HashingVectorizer followed by TruncatedSVD to obtain a dense latent semantic embedding of dimension 256. The combination of our models can be grouped into two categories. First are the embedding-based classifications: Embedding + Logistic Regression, Embed-

Method	AUC	Accuracy	F1
Embedding-based models			
Text-only + MLP	0.903 ± 0.030	0.825 ± 0.037	0.872 ± 0.032
FedMat + MLP	**0.911 ± 0.030**	**0.848 ± 0.034**	**0.889 ± 0.028**
Text-only + LinearSVM	0.892 ± 0.034	**0.840 ± 0.037**	**0.884 ± 0.029**
FedMat + LinearSVM	**0.902 ± 0.032**	0.830 ± 0.037	0.875 ± 0.030
Text-only + LogReg	0.891 ± 0.034	**0.835 ± 0.036**	**0.880 ± 0.029**
FedMat + LogReg	**0.908 ± 0.032**	0.832 ± 0.037	0.876 ± 0.031
Text-only + XGBoost	0.836 ± 0.041	0.784 ± 0.040	0.842 ± 0.033
FedMat + XGBoost	**0.913 ± 0.032**	**0.863 ± 0.035**	**0.903 ± 0.027**
Transformer-based models			
Text-only + DeBERTa	0.9981 ± 0.0062	0.9797 ± 0.0089	**0.9852 ± 0.0075**
FedMat + DeBERTa	**0.9993 ± 0.0054**	0.9797 ± 0.0085	0.9851 ± 0.0072
Text-only + GPT-2	**0.9783 ± 0.0126**	0.9340 ± 0.0139	0.9506 ± 0.0112
FedMat + GPT-2	0.9697 ± 0.0134	**0.9365 ± 0.0131**	**0.9529 ± 0.0108**

Table 1: Experiment results comparing Text-only and FedMat

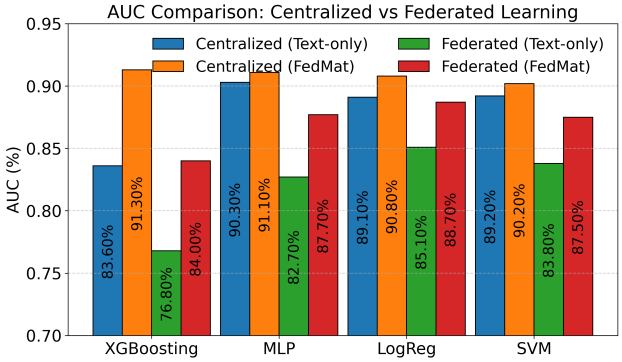

Figure 2: AUC Comparison: Centralized vs Federated Learning

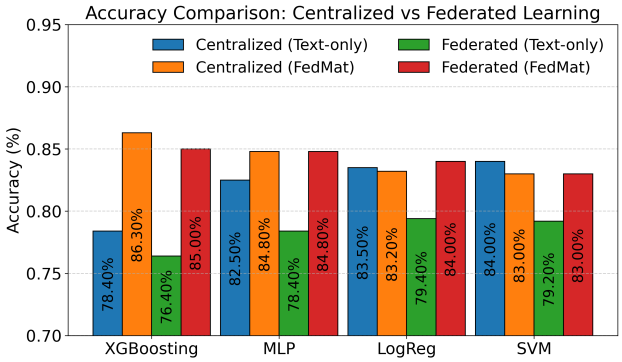

Figure 3: Accuracy Comparison: Centralized vs Federated Learning

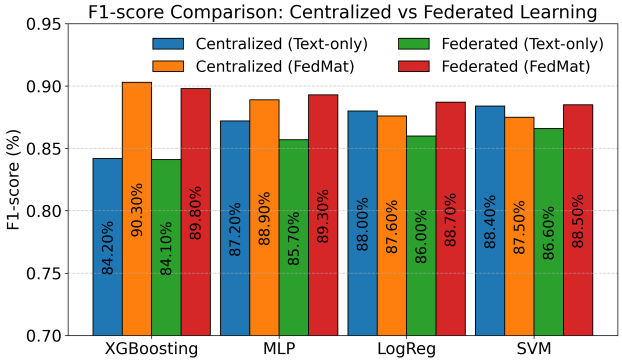

Figure 4: F1 Score Comparison: Centralized vs Federated Learning

ding + Linear SVM with calibrated probabilities, Embedding + XGBoost and Embedding + Multi-Layer Perceptron (MLP). The second category are the Transformer fine-tuning models: DeBERTa, and GPT-2. For multimodal transformer settings, we adopt a late fusion strategy by concatenating the transformer sentence representation with the materiality feature vector before the classification head. In order to solve the imbalanced data problem, we apply Synthetic Minority Over-sampling Technique (SMOTE) on the training dataset. In contrast, SMOTE is not applied to transformer fine-tuning; instead, we use a class-weighted loss to address label imbalance. We report Accuracy, F1-score, and AUC. For every experiment group, we additionally save the test predictions and predicted probabilities for reproducibility.

An analysis of the data set shows that materiality-related signals are common: 53.3% of sentences contain quantitative indicators, 33.0% include financial event cues, 48.8% exhibit contextual linguistic signals, and overall 59.2% contain at least one materiality component, supporting the suitability of the corpus for materiality-aware modeling.

Federated Learning Layer: Since the primary goal of the federated learning experiments is to assess the security and privacy benefits rather than to exhaustively compare all model variants in the federated setting, we only conduct federated evaluations on the embedding model in the centralized and federated learning experiments. This design choice significantly reduces the computational cost while preserving the validity of the security analysis, as the selected model represents a strong and competitive baseline. Specifically, we simulate a federated setting by partitioning the training data into K clients using a Dirichlet non-IID split; each client trains its local FedMat XGBoost classifier without sharing raw data, and global inference is obtained by aggregating client predictive distributions via data-size weighted averaging. Federated security performance is reported using macro-AUC (OVR).

Results and Discussion

Materiality Layer: We first test the embedding based classifications. As Table 1 shows, embedding-based classification performance under two settings, Text-only and FedMat. Overall, adding the material feature channel leads to consistent improvements in AUC across all models and supports that materiality rules provide complementary information beyond textual itself and help separate positive and negative samples more reliably across decision thresholds. This improvement can be observed significantly for XGBoost, where the FedMat setting increases the AUC from 0.836 to

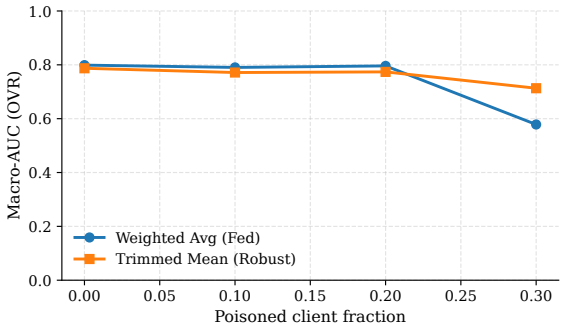

Figure 5: Macro-AUC under label-flipping poisoning: weighted averaging vs. trimmed-mean aggregation

0.913, and also yields clear gains in Accuracy from 0.784 to 0.863 and F1 from 0.842 to 0.903. Similar but smaller improvements are observed for MLP. For linear models such as LinearSVM and LogReg, FedMat mainly improves AUC, while accuracy and F1 decrease slightly. This pattern indicates that the added material features strengthen ranking quality, but may also change the optimal decision boundary or calibration, leading to small differences in threshold-dependent metrics. Therefore, for the embedding based classification algorithms, the material feature channel improve the financial document classification, especially beneficial for the non-linear classifiers. This suggests that FedMat is particularly valuable when text-only models still have room for improvement, highlighting its role as a complementary signal rather than a replacement for strong language models. The results of the transformer experiments in Table 1 also show the trend that FedMat achieves better results. Although a small centralized–federated performance gap is expected, future work may explore personalized federated learning or improved aggregation strategies to further narrow this gap in real-world deployments.

Federated Learning Layer: Figures 2, 3, 4 show that in all three metrics (AUC, Accuracy, and F1), the results consistently show that incorporating materiality features improves performance in both centralized and federated learning. In the centralized setting, FedMat produces the strongest overall results, with AUC reaching 90.2%–91.3%, compared to 83.6%–90.3% for Text-only, and the largest AUC gain occurs for XGBoost, which improved 7.7%, from 83.6% to 91.3%. The same trend holds for Accuracy (from 76.4% to 79.4% to 83.0% to 85.0%) and F1 (from 84.1% to 86.6% to 88.5% to 89.8%), demonstrating that materiality variables provide robust complementary signals and help mitigate performance loss in privacy-preserving federated deployment.

The figure 5 reports the robustness of the federated FedMat XGBoost model under a label-flipping poisoning attack by varying the fraction of malicious clients from 0% to 30% and tracking OVR on a fixed holdout test set. In this setting, a selected subset of clients flips sentiment labels during local training, injecting systematic noise, and biases local decision boundaries. The results show that both aggregation rules are stable when the poisoned fraction is small (0–20%), with macro-AUC remaining around 0.78 to 0.80, indicating that the federated ensemble is largely dominated by the majority benign clients in this regime. However, when the poisoned fraction reaches 30%, standard data-size weighted averaging becomes substantially more vulnerable, with macro-AUC dropping sharply to approximately 0.58, suggesting a threshold effect where corrupted client outputs begin to meaningfully distort the aggregated predictive distribution. In contrast, robust trimmed-mean aggregation degrades more gracefully and maintains a substantially higher macro-AUC (approximately 0.71) at 30% of poisoning, consistent with its ability to attenuate extreme or outlying client predictions. Overall, the figure demonstrates that while federated learning can retain strong utility under mild adversarial presence, simple robust aggregation provides a practical defense that improves resilience under stronger poisoning conditions.

Conclusion

Our study introduces FedMat, a federated multimodal framework for trustworthy financial document analysis that addresses key limitations of traditional approaches, namely the reliance on text-only sentiment signals and the risk of exposing sensitive financial data. By integrating quantitative financial indicators, categorical event features, and contextual linguistic cues, FedMat generates sentence-level materiality scores that enhance the economic interpretability of textual disclosures. Experimental results demonstrate that incorporating materiality-aware features consistently improves predictive performance across both embedding-based and transformer-based models. The gains are particularly pronounced for nonlinear classifiers; for example, XGBoost models configured with FedMat achieve substantial improvements in AUC, accuracy, and F1-score. When combined with federated learning, the framework enables multiple financial institutions to collaboratively train models without sharing raw data, thus strengthening privacy protection while improving generalization across decentralized datasets. Our experiments further show that the federated FedMat setting maintains strong predictive performance even under adversarial conditions, and that robust aggregation strategies such as trimmed-mean effectively mitigate the impact of label-flipping attacks. Overall, the results indicate that the integration of multimodal representations that are materiality-sensitive with federated learning improves both the reliability and the interpretability of financial sentiment analysis while supporting secure real-world deployment.

From a business perspective, FedMat offers a practical pathway for organizations to deploy collaborative AI models while complying with data governance and privacy regulations. Financial institutions could integrate the framework within existing reporting analytics pipelines, using local feature extraction and federated aggregation to share learning signals without exposing proprietary data. Key implementation challenges include harmonizing reporting standards across institutions, aligning materiality definitions, and managing communication overhead in distributed environments.

Addressing these organizational and operational considerations will be essential for translating the proposed framework into scalable enterprise applications, which we identify as an important direction for future work.

References

Alhasawi, Y.; Almtrf, A. A.; and Asad, M. 2025. A Federated Approach to Scalable and Trustworthy Financial Fraud Detection. *Security and Privacy*, 8(5): e70099.

Aljunaid, S. K.; Almheiri, S. J.; Dawood, H.; and Khan, M. A. 2025. Secure and transparent banking: explainable AI-driven federated learning model for financial fraud detection. *Journal of Risk and Financial Management*, 18(4): 179.

Andrade, G.; Mitchell, M.; and Stafford, E. 2001. New evidence and perspectives on mergers. *Journal of economic perspectives*, 15(2): 103–120.

Campbell, J. L.; Chen, H.; Dhaliwal, D. S.; Lu, H.-m.; and Steele, L. B. 2014. The information content of mandatory risk factor disclosures in corporate filings. *Review of Accounting Studies*, 19(1): 396–455.

Capital One Financial Corporation. 2019. Capital One Announces Data Security Incident. https://www.capitalone.com/about/newsroom/capital-one-announces-data-security-incident/. Accessed: 2025-11-18.

Cho, S.; Moon, J.; Bae, J.; Kang, J.; and Lee, S. 2023. A framework for understanding unstructured financial documents using RPA and multimodal approach. *Electronics*, 12(4): 939.

Dias, D.; Borbinha, J.; and Mendonça, P. 2022. Lessons from the Equifax and Capital One Data Breaches on Social Amplification of Risk. *Review of Integrative Business and Economics Research*, 11(4): 71–90.

Edmans, A.; Pu, D.; Zhang, C.; and Li, L. 2024. Employee satisfaction, labor market flexibility, and stock returns around the world. *Management Science*, 70(7): 4357–4380.

Fang, K. C.; and Jacobs, B. 2000. Clarifying and protecting materiality standards in financial statements: A review of SEC Staff Accounting Bulletin 99. *The Business Lawyer*, 1039–1064.

Federal Trade Commission (FTC). 2024. Equifax Data Breach Settlement. https://www.ftc.gov/enforcement/refunds/equifax-data-breach-settlement. Accessed: 2025-11-18.

Frank, M. Z.; and Goyal, V. K. 2003. Capital structure decisions. *Available at SSRN 396020*.

Grewal, J.; Hauptmann, C.; and Serafeim, G. 2021. Material sustainability information and stock price informativeness. *Journal of Business Ethics*, 513–544.

Jegadeesh, N.; and Livnat, J. 2006. Revenue surprises and stock returns. *Journal of Accounting and Economics*, 41(1-2): 147–171.

Khan, S.; Kabanov, I.; Hua, Y.; and Madnick, S. 2022. A systematic analysis of the capital one data breach: Critical lessons learned. *ACM Transactions on Privacy and Security*, 26(1): 1–29.

Kinney, W.; Burgstahler, D.; and Martin, R. 2002. Earnings surprise "materiality" as measured by stock returns. *Journal of Accounting Research*, 40(5): 1297–1329.

Kothari, S. P. 2001. Capital markets research in accounting. *Journal of accounting and economics*, 31(1-3): 105–231.

Langevoort, D. C. 1998. Commentary: Stakeholder Values, Disclosure, and Materiality. *Cath. UL Rev.*, 48: 93.

Liao, C.-F. 2025. ESG Disclosure Frequency and Its Association with Market Performance: Evidence from Taiwan. *Sustainability*, 17(17): 7812.

Liu, L.; and Han, M. 2024. Data sharing and exchanging with incentive and optimization: a survey. *Discover Data*, 2(1): 2.

Liu, L.; Kong, D.; Chen, Y.; Pouriyeh, S.; Zhou, Y.; and Han, M. 2025. Trust in Stablecoins: A Survey from AI-Driven Perspective. *Available at SSRN 5619876*.

Liu, L.; Kong, Y.; Li, G.; and Han, M. 2023. Fairshare: an incentive-based fairness-aware data sharing framework for federated learning. In *International conference on intelligent robotics and applications*, 115–126. Springer.

Loughran, T.; and McDonald, B. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of finance*, 66(1): 35–65.

Malo, P.; Sinha, A.; Korhonen, P.; Wallenius, J.; and 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.

Mosaiyebzadeh, F.; Pouriyeh, S.; Han, M.; Liu, L.; Xie, Y.; Zhao, L.; and Batista, D. M. 2024. Privacy-preserving federated learning-based intrusion detection system for ioht devices. *Electronics*, 14(1): 67.

Mosca, C.; and Picciau, C. 2020. Making non-financial information count: accountability and materiality in sustainability reporting. *Finance Durable et Droit: Perspectives Comparées (Hugues Bouthinon-Dumas, Bénédicte François & Anne-Catherine Muller eds., 2020, Forthcoming), Bocconi Legal Studies Research Paper*, (3536460).

Pappula, K. K.; and Rusum, G. P. 2023. Multi-Modal AI for Structured Data Extraction from Documents. *International Journal of Emerging Research in Engineering and Technology*, 4(3): 75–86.

Pizzi, S.; Principale, S.; and De Nuccio, E. 2023. Material sustainability information and reporting standards. Exploring the differences between GRI and SASB. *Meditari accountancy research*, 31(6): 1654–1674.

Sousa, A. R. C.; Pacheco, J.; and Rover, S. 2024. Disclosure and materiality of intangible assets in the value relevance of the brazilian stock market. *Enfoque: Reflexão Contábil*, 43(2): 151–173.

Tang, Y.; and Liang, Y. 2024. Credit card fraud detection based on federated graph learning. *Expert Systems with Applications*, 256: 124979.

van der Lugt, C. T.; Bakker, H.-P.; and Mans-Kemp, N. 2025. Materiality in reporting integration in South Africa: A natural language processing analysis. *South African Journal of Economic and Management Sciences*, 28(1): 5717.