FSE 570 Capstone Project Proposal
Earnings Call and Risk Intelligence Engine for Financial Decision Support

Shalin N. Bhavsar sbhavsa8@asu.edu
Ramanuja M. A. Krishna rkrish79@asu.edu
Freya H. Shah fshah14@asu.edu
Kruthika Suresh ksures21@asu.edu
Harihara Y. Veldanda hveldan1@asu.edu

Arizona State University

1. Problem Statement

Public companies in the United States conduct approximately 10,000 earnings calls annually, producing transcripts ranging from 8,000 to 15,000 words. These transcripts contain forward-looking statements, risk disclosures, and management language patterns critical for evaluating earnings quality and downside risk, yet their unstructured and lengthy nature makes systematic analysis infeasible using manual review. Current industry practice relies on manual reading, basic keyword searches, or delayed market reactions. Manual analysis requires hours to days per earnings cycle and does not scale, causing early warning signals of earnings manip- ulation or deteriorating fundamentals to be identified only after delayed market reactions and short-term mispricing following earnings releases. This project develops an Earnings Call and Risk Intelligence Engine that automatically extracts sentiment, linguistic uncertainty, and risk signals from earnings calls and integrates them with structured financial data, reducing analysis time from days to minutes. This is a data-driven engineering problem because it requires the design of scalable pipelines to ingest, preprocess, and integrate large volumes of heterogeneous unstructured text and structured financial time-series data, followed by the deployment of validated machine learning models under strict temporal constraints. The system must satisfy the following requirements:

- Scalable processing of 10,000+ earnings call transcripts annually.
- Accurate and interpretable extraction of sentiment, risk, and uncertainty signals.
- Integration of unstructured text with structured financial time-series data.
- Predictive outputs supporting timely post-earnings decisions.

Primary stakeholders include institutional investors, equity analysts, portfolio managers, and corporate risk teams. At a broader level, improved interpretation of earnings disclosures can reduce mispricing, improve capital allocation efficiency, and contribute to greater market stability during periods of economic uncertainty.

2. Data Sources

The project uses multiple large, heterogeneous, and publicly accessible datasets:

- **Earnings Call Transcripts**: 13,500 transcripts from Seeking Alpha and SEC EDGAR (2018-2023, S&P 500 companies) with metadata (ticker, date, quarter) and structured content (speaker roles, prepared remarks, Q&A sections).
- **Financial and Market Data**: Quarterly statements from SEC EDGAR with 28 features (revenue, EPS, margins, leverage, cash flow). Daily stock data from Yahoo Finance (OHLC, volume, adjusted close accounting for splits/dividends) aligned by ticker and date.

Key challenges include transcript cleaning, speaker segmentation, temporal alignment between text and financial time-series data, and handling missing or noisy records. Potential sourcing challenges include API rate limits, inconsistent HTML structures across transcripts, and incomplete historical coverage for smaller-cap firms. All datasets are public and contain no personally identifiable information.

3. Methodology

- **Prediction Target**: Binary classification of 3-day post-earnings abnormal returns (market-adjusted: stock return minus S&P 500 return). Three-day window captures post-earnings drift (Bernard and Thomas, 1989) while minimizing unrelated noise.

- **Text Processing**: Transcript normalization, tokenization, lemmatization, and speaker-role identification (management vs. analysts).
- **Feature Extraction**: FinBERT generates sentiment scores and 768-dim embeddings (Financial PhraseBank F1: 0.97 vs. BERT 0.89; publicly available; 110M parameters for deployment). Loughran-McDonald dictionary quantifies linguistic uncertainty. Management-analyst sentiment divergence captures Q&A tone shifts.
- **Data Integration**: Early fusion concatenates mean-pooled FinBERT embeddings (768-dim) with normalized quarterly financials (28 features) forming 796-dim vectors per event, enabling sentiment-fundamental interaction learning. Temporal alignment prevents lookahead bias.
- **Predictive Modeling**: Logistic regression baseline. XGBoost primary model (gradient boosting for non-linear interactions, native missing value handling, SHAP interpretability). Hyperparameters via grid search (learning rate: [0.01, 0.1], depth: [4, 6, 8], trees: [50, 100, 150]). Random forest validates architecture-independence. SMOTE balances positive/negative classes.
- **Validation:** 70/30 chronological split with 5-fold time-series CV maintaining temporal order. Metrics: accuracy, F1, ROC-AUC. Baselines: sentiment-only, financial-only. Success: ≥3% ROC-AUC improvement (statistically significant via DeLong test, $p < 0.05$).

4. Project Management: Timeline and Risks

Week	Milestone	Monitoring Criteria
1	Finalize scope, datasets, and evaluation metrics	Scope approved, API access verified
2	Collect earnings calls and financial data	≥90% data collected
3	Transcript cleaning and normalization	Parsing pipeline operational
4	Speaker segmentation and temporal alignment	≥85% transcripts parsed
5	Baseline financial-only modeling	Baseline ROC-AUC ≥0.55 established
6	NLP feature extraction (FinBERT)	FinBERT embeddings for 100% of transcripts
7	NLP-enhanced model training	≥3% ROC-AUC over baseline
8	Model validation and comparison	Cross-validation stability
9	Error analysis and refinement	Sensitivity metrics reviewed
10	Final report and presentation	All deliverables completed

Status Check-ins:

- **Status Update 1 (Week 4)**: Cleaned dataset with ≥85% parsing success, baseline model with documented ROC-AUC ≥0.55, identified data quality issues with resolution plan.
- **Status Update 2 (Week 8)**: Integrated models achieving ≥3% ROC-AUC improvement over baseline, completed cross-validation, documented error patterns and model limitations.

Key Risks and Mitigation:

- **Data Quality Risk**: Monitored weekly via automated parsing logs tracking success rate by source. If rate drops below 80% in any source, the designated data engineer will within 48 hours implement source-specific parsing rules or exclude that source. Contingency: Pre-identified backup transcript sources (FactSet, Capital IQ) if primary sources fail.
- **Model Performance Risk**: Tracked after each training run via ROC-AUC delta from baseline. If improvement is <2% by end of Week 7, team lead will convene model review to implement pre-defined fallback sequence: (1) SHAP-based feature pruning, (2) switch to LightGBM with pre-tuned hyperparameters, (3) ensemble baseline models if all else fails.
- **Resource Constraint Risk**: Monitored via processing time logs per batch. If any batch exceeds 8 hours per 1,000-transcript batch, team will immediately switch to stratified sampling (reduce to 9,000 transcripts maintaining sector representation) or switch from FinBERT to lightweight DistilFinBERT to ensure Week 6 deadline is met.