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 manipulation 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**: Approximately 12,000–15,000 transcripts from Seeking Alpha and SEC EDGAR, available in text or HTML format, spanning multiple years and U.S.-listed firms. Each transcript is structured with metadata fields (company ticker, date, fiscal quarter) and content fields (speaker identifiers, role labels, utterance text, Q&A segmentation markers). Transcripts include identifiable speaker roles (executives vs. analysts), prepared remarks, and Q&A sections, with lengths ranging from 8,000 to 15,000 words.
- **Financial and Market Data**: Quarterly financial statements from SEC EDGAR structured as tabular records with fields including revenue, net income, earnings per share, operating margins, leverage ratios, and cash flow metrics. Daily stock data from Yahoo Finance includes open, close, high, low, volume, and adjusted close prices, all aligned with earnings announcement dates via ticker and timestamp matching.

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

- **Text Processing**: Transcript normalization, tokenization, lemmatization, and speaker-role identification to preserve contextual differences between management statements and analyst questions.
- **Feature Extraction**: FinBERT, pretrained on financial-domain text, is used to generate sentiment scores, linguistic uncertainty measures, and contextual embeddings. Financial-domain pretraining enables more accurate interpretation of earnings-specific language, hedging, and risk disclosures than generic language models. Topic modeling is applied to identify recurring risk-related themes.
- **Data Integration**: Text-derived embeddings and sentiment features are concatenated with aligned quarterly financial indicators to form a unified feature vector for each earnings event, enabling joint modeling of linguistic signals and numerical fundamentals.
- **Predictive Modeling**: Logistic regression serves as baseline. Random forests and XGBoost are selected to capture non-linear interactions while maintaining interpretability, with tree-based models chosen for robustness on tabular data with mixed feature types.
- **Model Validation**: An expanding-window temporal train-test split is used to prevent data leakage, with 5-fold cross-validation applied on the training set. Performance is evaluated using accuracy, F1-score, ROC-AUC, and regression error metrics, and compared against sentiment-only and financial-only baseline models.

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	$\geq$90% data collected
3	Transcript cleaning and normalization	Parsing pipeline operational
4	Speaker segmentation and temporal alignment	$\geq$85% transcripts parsed
5	Baseline financial-only modeling	Baseline ROC-AUC established
6	NLP feature extraction (FinBERT)	Features generated successfully
7	NLP-enhanced model training	$\geq$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 $\geq$85% parsing success, baseline model with documented ROC-AUC benchmark, identified data quality issues with resolution plan.
- **Status Update 2 (Week 8)**: Integrated models achieving $\geq$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 24 hours, team will immediately switch to stratified sampling (reduce to 8,000 transcripts maintaining sector representation) or switch from FinBERT to lightweight DistilFinBERT to ensure Week 6 deadline is met.