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

In the United States, S&P 500 companies hold four earnings calls per year, generating tens of thousands of transcripts across all public companies. These 7,000-15,000 word transcripts contain critical information like forward-looking statements, management tone, and risk indicators essential for assessing earnings quality and downside risk. However, their length and unstructured nature make large-scale analysis difficult. Current methods rely heavily on manual work or basic keyword searches, limiting scalability and delaying the detection of early warning signs about earnings manipulation or rising uncertainty.

This project proposes an Earnings Call and Risk Intelligence Engine that automatically extracts sentiment, uncertainty, and risk-related linguistic features from transcripts and integrates them with structured financial data, reducing analysis time from hours to minutes. The system features an Agentic AI Orchestration Layer, implemented using LangChain and LangGraph, that automates the entire workflow—transcript ingestion, cleaning, feature extraction, model execution, validation, and reporting—end-to-end without human intervention. The system must satisfy: (1) scalable processing of thousands of transcripts per year, (2) accurate extraction of sentiment, risk, and uncertainty signals, (3) robust integration of unstructured text with structured financial time-series data, (4) automated orchestration via an Agentic AI framework, and (5) timely predictive outputs for post-earnings decisions. Primary stakeholders include institutional investors, equity analysts, portfolio managers, and risk management teams.

2. Data Sources

Among the data sets used are many large, heterogeneous, and publicly available:

- **Earnings Call Transcripts:** There are 13,500 transcripts available from Seeking Alpha (seekingalpha.com/earnings) and SEC EDGAR (sec.gov/edgar) (2018-2023, S&P 500 companies). Each transcript is structured with metadata fields (ticker, date, quarter) and content fields (speaker roles, prepared remarks, Q&A sections, utterance text), with lengths ranging from 8,000 to 15,000 words. These transcripts provide the unstructured text needed for extracting sentiment, uncertainty, and risk signals critical to the prediction task.
- **Financial and Market Data**: Quarterly financial statement data reported on SEC EDGAR and formatted as tabular records with 28 attributes or features, such as revenue, EPS, net income, operating margins, leverage, and cash flows. Daily stock information from Yahoo Finance (finance.yahoo.com) contains OHLC information, volume, and adjusted close prices normalized for stock splits and dividends, and can be matched on stock and date.

Key challenges include building a system capable of developing, designing, obtaining transcripts, segmenting speakers, performing temporal alignment between text and financial time-series data, and handling missing/noisy records. Other sourcing challenges - API rate limits, inconsistent structures of HTMLs across transcripts, and incomplete historical coverage for smaller-cap firms. All datasets are publicly available with no personally identifiable information.

3. Methodology

- **Prediction Target:** Prediction Target: Binary classification of abnormal returns 3 days after the earnings announcement, where abnormal returns = returns - S&P 500 returns. This time period is long enough to capture the post-earnings announcement drift (Bernard and Thomas, 1989), but short enough to exclude other types of noise.
- **Text Processing**: Transcript Normalization, Tokenization, Lemmatization, and Speaker-Roles (Management vs. Analysts).
- **Feature Extraction**: FinBERT produces sentiment scores and embedding vectors with 768 dimensions, as well as sentiment scores. FPhraseBank F1: 0.97 vs. BERT: 0.89; publicly available; 110M parameters. Loughran–McDonald dictionary quantifies uncertainty in language. Management Analyst Sentiment Divergence

- **Data Integration**: The early fusion integrates the concatenated mean pooling of FincBERT, with its 768 dimensions, and normalizes quarterly information with 28 features into a combined 796-dimensional vector space. The temporal aspect prevents lookahead bias.
- **Predictive Modeling**:Logistic regression baseline. XGBoost as primary model (handles nonlinear relationships, missing data; SHAP for interpretability). Hyperparameters tuned via grid search (learning rate: [0.01, 0.1], depth: [4, 6, 8], trees: [50, 100, 150]). Random forest for model-architecture-independence verification.
- **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: $\geq$3% ROC-AUC improvement (statistically significant via DeLong test, $p < 0.05$).
- **Workflow Orchestration and Pipeline Control:** A controlled workflow built with LangGraph manages the end-to-end pipeline (ingestion $\rightarrow$ preprocessing $\rightarrow$ fusion $\rightarrow$ modeling $\rightarrow$ validation $\rightarrow$ reporting). The system automatically checks transcript completeness, speaker segmentation, and temporal alignment to prevent lookahead bias. If parsing fails beyond a set threshold, it falls back to predefined feature sets (e.g., financial-only baseline). All steps are logged to ensure reproducibility, and performance is evaluated through latency and failure-recovery testing against a non-orchestrated pipeline.

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 $\geq$0.55 established
6	NLP feature extraction (FinBERT)	FinBERT embeddings for 100% of transcripts
7	NLP-enhanced model training	$\geq$3% ROC-AUC over baseline
8	Model validation and comparison	Cross-validation stability
9	Agentic orchestration integration	Automated pipeline runs end-to-end
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 $\geq$0.55, data quality issues identified with resolution plan.
- **Status Update 2 (Week 8)**: Developed integrated models with $\geq$3% ROC AUC improvement over the baseline, cross-validation of the models done, error patterns and model limitations documented.

Key Risks and Mitigation:

- **Data Quality Risk**: Tracked on a weekly basis using automated parsing log data, which tracks the success rate by source. If any one of these sources has a success rate less than 80%, the associated data engineer shall apply source-specific parsing rules within 48 hours. Contingency: Backup sources of transcripts - FactSet, Capital IQ shall be in place in case primary sources have issues.
- **Model Performance Risk**: Should be monitored after every training event based on the ROC AUC Delta calculation from the base metric. If this value is less than 2% at the end of Week 7, the leader must call a model review to enact a predetermined set of fallbacks in sequence: (1) feature pruning using SHAP values, (2) switching the model to LightGBM with pre-selected hyperparameters, and (3) an ensemble baseline model if all else fails.
- **Resource Constraint Risk**: This is monitored by tracking processing time per batch. In case a batch crosses the limit of 8 hours per 1,000 transcripts, it will be immediately switched to stratified sampling, reducing it to 9,000 while maintaining representation in the sectors, or switching to the lighter version of FinBERT, known as DistilFinBERT, to ensure that it meets in time the Week 6 deadline.
- **Automation Risk:** Agentic orchestration logs all pipeline steps and validation checkpoints. Any failure in sequencing, validation thresholds, or temporal checks triggers automatic rollback to the last verified state.