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

Every public company in the United States hosts 10,000 earnings calls per annum, which generate transcripts ranging from 8,000 to 15,000 words. These transcripts include forward-looking statements, risk disclosures, and language patterns critically important for assessing earnings quality and downside risk. Yet, their unstructured and lengthy nature makes systematic analysis infeasible using manual review. The existing practice followed is manual reading, fundamental keyword search, or lagged market response. Manual analysis is a time-consuming process that can take hours or days for each earnings report. Furthermore, such a cycle does not scale, thereby creating early signals of suspected earnings manipulations or weakening fundamentals, and thus can only be identified in terms of their delayed market reactions and short-run mispricing associated with earnings announcement episodes. This project develops an Earnings Call and Risk Intelligence Engine that automatically infers sentiment, linguistic uncertainty, and incorporates signals from earnings calls with structured financial data, cutting analysis times from days to minutes. This problem qualifies as a data-driven engineering issue since it demands the development of generalized pipelines for ingest, preprocess, and integrate large volumes of heterogeneous unstructured text and structured financial time series data, followed by the application of well-developed machine learning models with tight temporal constraints.
The system must meet the following criteria:

- Scalable processing of over 10,000 earnings call transcripts every year.
- Extraction of sentiment, risk, and uncertainty signals both accurately and in an interpretable manner.
- The integration of unstructured text with structured financial time-series data.
- Predictive outputs to aid in timely post-earnings decisions.

Primary stakeholders include institutional investors, equity analysts, portfolio managers, and corporate risk teams. Improved interpretation of earnings disclosures might also, at a broader level, reduce the incidence of mispricing, improve capital allocation efficiency, and add to the market stability in periods of economic uncertainty.

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 cleaning the 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 reference model. XGBoost primary model (gradient boosting for nonlinear relationships, missing data handling, SHAP interpretation method). Hyperparameters tuned via grid search (learning rate: [0.01, 0.1], depth: [4, 6, 8], trees: [50, 100, 150]). Random forest for verification of the effectiveness of the model architectural independence.
- **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, $p < 0.05$).

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	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 $\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 in order 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.