Converting Expert Deliberation into Financial Signals Through A Context-Aware NLP Pipeline

Vivek Batra
Franklin Templeton Investments
Mumbai, India
vivek.batra@franklintempleton.com

Kristin Chen
Blend360
Columbia, USA
kristin.chen@blend360.com

Sanjiv R. Das
Santa Clara University
Santa Clara, USA
srdas@scu.edu

Samuel Judge
Franklin Templeton Investments
Boston, USA
samuel.judge@franklintempleton.com

Harshad Khadilkar
Franklin Templeton Investments
Mumbai, India
harshad.khadilkar@franklintempleton.com

Sukrit Mittal
Franklin Templeton Investments
Hyderabad, India
sukrit.mittal@franklintempleton.com

Amir Nasrollahzadeh
Blend360
Boston, USA
amir.nasrollahzadeh@blend360.com

Daniel Ostrov
Santa Clara University
Santa Clara, USA
dostrov@scu.edu

Jacob Sisk
Franklin Templeton Investments
San Ramon, USA
jacob.sisk@franklintempleton.com

***Abstract*—We introduce the CDSP (context-conditional deliberation signal) pipeline, converting an investment committee's meeting transcripts into structured predictive features. CDSP segments the meeting transcripts into topical chunks, assigns asset-class context labels using a large language model (LLM), maps financial keywords to a pre-determined taxonomy of labels, and constructs complementary features: sentiment polarity and mention frequency. This feature engineering framework is applied to a dataset spanning 48 monthly committee meetings, to predict if global equities will perform better or worse than global bonds in the following month. In experiments with engineered features, raw transcript text, sentence embeddings, and combined representations, the prediction accuracy ranges from 62% to 73%, compared to always choosing stocks, which outperforms bonds 60.4% of the time. The best (73% accurate) model combines sentence embeddings with engineered CDSP features, achieving a 0.73 F1 score (although this is not statistically significant compared to always choosing stocks). Sentiment carries a stronger signal than mention frequency for several taxonomy categories. These findings suggest that experts' deliberations may contain forward-looking information that context-aware NLP can extract.**

***Index Terms*—NLP, portfolio management, text classification, sentence embeddings, sentiment analysis, topical analysis**

I. Introduction

Firms rarely make high-stakes decisions by simply observing numbers and acting on them. The underlying subjectivity, especially, the debate over upside versus downside, the comparison of alternative scenarios, helps them form a collective view on the best path forward. Investment committees or portfolio research groups operate in a similar manner on a regular basis. They interpret uncertain information, debate future possibilities, and decide how to position themselves before that future is known. These discussions are often recorded for review, compliance and institutional memory, effectively creating an archive of expert reasoning. With the availability of Large Language Model (LLM)-based tools, there is now an opportunity to systematically extract the thought process behind those decisions: for example, how a geopolitical condition was weighed against an offsetting macroeconomic condition, and how well the resulting decision ultimately performed.

At first glance, it may seem straightforward to feed such historical archives into LLMs with large context windows, and query them on an as-needed basis. However, this raises questions that are particularly relevant in context-sensitive investment applications:

- Do investment committee transcripts contain structured predictive signals about market performance in the short term?
- Does the meaning and the prediction power of financial language change depending on where and in which context it appears within the discussion?

Although the LLM may provide answers to both questions, uncertainty remains because LLMs are largely black boxes. This paper attempts to address that uncertainty by treating the historical archive as a structured data source and evaluating its empirical merit in making predictions. We study whether the language used in investment committee meetings contains structured signals about the direction of equities versus bonds in the following month. We assess whether modern natural language processing (NLP) can transform expert discussions into interpretable features that can be used in a supervised learning model.

In the context of this paper, an investment committee meets

each month to discuss macroeconomic conditions, outlooks for asset classes like equities and bonds, and portfolio positioning across these assets. The committee's objective is to form a forward-looking view about relative opportunities and risks. For example, a committee may discuss whether equity valuations are attractive relative to bond yields, or whether macroeconomic risks justify reducing exposure to riskier assets. These conversations therefore contain opinions that are guided by perceived financial patterns.

Our focus is the simple prediction problem of whether or not equities will outperform bonds in the following month. This captures allocation decision information in a simple form. The key questions here are whether the committee's discussion contains enough information to help predict this, and, more importantly, whether NLP techniques can use this information to generate meaningful predictions.

To answer these key questions, we introduce the Context-conditional Deliberation Signal Pipeline (CDSP). CDSP cleans and segments meeting transcripts into topic-based chunks. It then uses an LLM to assign to each chunk one or more context labels from the set $\mathcal{C}$ ={equity, bonds, cross-asset, macroeconomic-outlook, meeting-logistic, other}. From a vocabulary of 2416-keyword domains, financial keywords are extracted from each chunk and mapped to a curated taxonomy of 15 subcategories contained within four investment attribute categories (given in Table I). For each topic and context label, we construct two feature categories: *sentiment*, which measures how positive or negative the surrounding language is, and *mention frequency or density*, which measures how much attention the committee gives to that topic. The key idea is *context-conditioning*. The same financial phrase may have a different meaning depending on where it appears in the meeting. For example, discussion of valuation or credit conditions during an equity-focused segment can carry a different signal than the same language during a macroeconomic overview.

The existing literature on text-based financial prediction operates largely at the document level. For example, foundational work demonstrated that negative media tone predicts stock price declines [1], finance-specific sentiment in 10-K filings predicts filing-day returns [2], and earnings call tone changes are informative about future fundamentals and investor/analyst reactions [3]. More recent work has extended these ideas to structured representations using neural language models [4], [5] and LLMs [6]. A consistent limitation across this body of work is the treatment of text as a flat signal. [7] showed that deliberation patterns contain information about policy decisions, and Osowska et al. [8] empirically confirmed its link to markets' reaction. While text-based signal detection has been an extensive topic of study from the early days of financial NLP, using internal investment committee discussions comprises a novel source of signals [9].

In this paper, we compare three ways of representing the transcripts: (1) using engineered semantic features produced by CDSP, (2) using the transcript text directly, and (3) using sentence embeddings, which provide dense numerical representations. We also test whether combining engineered features with embeddings improves performance. Across these approaches, the classification accuracy for the next-month market movement ranges from 62%–73%, where the best results come from combining sentence embeddings with engineered CDSP features, achieving 73% accuracy and a 0.73 F1 score under repeated stratified cross-validation.

The remainder of the paper is structured as follows. Section II reviews related work. Section III describes the dataset. Section IV details the CDSP framework. Section V presents feature analysis. Section VI describes the predictive models and results. Section VII discusses implications and limitations. Section VIII concludes.

II. Related Work

A. Text-Based Prediction in Finance

NLP has been actively explored for financial applications such as sentiment analysis, financial forecasting, portfolio management, risk management, financial narrative processing, and explainable AI [9]. Most studies in this area explored using public or corporate text, including news, filings, annual reports, earnings calls, analyst reports, social media, and online forums (like X, Reddit, etc.). These sources have been shown to contain information relevant to returns, volatility, and investor reactions [3], [10], [11]. For example, [1] showed that media pessimism can predict downward pressure on stock prices and higher trading volume, while sentiment signals extracted from investor discussions and social media have also been linked to market direction [12], [13]. [2] showed that these signals can be improved using finance-specific dictionaries over general sentiment lexicons.

In recent years, these ideas have been extended using transformer-based models, especially LLMs. [4] adapted general-context BERT into FinBERT, specifically for financial sentiment analysis; [5] proposed BloombergGPT, demonstrating the value of finance-specific large-scale pretraining of LLMs; and [6] showed that the LLM-based sentiment extracted from financial headlines contains predictive information for near-term market returns. Notably, much of this literature treated text as document-level or event-level signals, and lesser attention has been given to internal structure of expert discussions.

B. Expert Communication and Committee Deliberation

Another stream of existing literature explores financial narratives. This includes earnings calls, analyst communications, and corporate reports, which usually contain qualitative information beyond numbers. [3] studied influence of changes in tone during earnings calls over analysts and investors. Some studies showed that linguistic cues in conference calls can help detect deceptive communication or future financial risk [14], [15]. These studies revealed that expert language can encode soft information, potentially relevant to financial outcomes.

The closest work to this paper comes from the analysis done for central bank communications. [7] used computational linguistics to study deliberation within the Federal Open Market Committee (FOMC). [8] showed that the topical and sentiment

features from the FOMC post-meeting statements help predict market reactions. These papers effectively demonstrated that committee language can contain meaningful signals worth exploring, however their focus has primarily been on public monetary-policy communications.

C. Portfolio Management and Soft Information

Portfolio-related NLP usually relies on external text sources such as news, social media, company descriptions, etc. Existing studies have examined how NLP-based forecasting and sentiment signals can support asset allocation decisions and portfolio construction [16], [17]. Some have used news-derived embeddings and semantic representation of firms to improve portfolio optimization [18], [19]. The more recent work of [20] also includes news sentiments into portfolio construction processes.

Some studies on soft-information emphasized the inclusion of qualitative and judgment based information that is difficult to encode in standard quantitative variables [21], [22]. The challenge of transforming deliberative information into structured features that are both predictive and interpretable still remains.

D. Structured and Context-Aware Representations

Textual data can be represented in different ways: using lexicons, topic models, neural embeddings, and LLM-derived features, etc. Topic models such as Latent Dirichlet Allocation identify latent themes in documents [23], while sentence-level embeddings provide dense semantic representation for classification and retrieval [24]. AutoML frameworks further provide the capability to combine tabular, textual, and embedding-based features efficiently [25], [26]. These methods are powerful, but they often reduce interpretability since the information of which financial concepts drove the prediction gets lost.

This paper addresses this limitation through CDSP, a context-conditioned pipeline for investment committee deliberations, allowing the model to preserve the structure of deliberation rather than treating the transcript as a flat document.

III. Data

A. Meeting Transcripts

In this paper, the data set is composed of transcripts of investment committee meetings usually convened on a monthly basis between June 2020 and February 2025. The committee actively discussed and deliberated macroeconomic conditions, specific asset-level outlooks for equities and bonds, and relative portfolio positioning (equities versus bonds). The meetings took place via video conference, hence transcripts created using a third-party software were made available. The dataset contained 48 transcripts due to uneven annual coverage. For example, year 2022 is over-represented (10 of 48 meetings) since the committee convened more frequently, likely due to post-COVID inflation.

B. Target Variable

In this paper, the prediction target is the *binary market direction* for the month following each meeting. We define this binary market direction by whether or not global equities outperformed bonds in the subsequent calendar month:

$$y_t = \begin{cases} 0 & \text{if } R^{\text{eq}}_{t+1} \leq R^{\text{bond}}_{t+1} \\ 1 & \text{if } R^{\text{eq}}_{t+1} > R^{\text{bond}}_{t+1}, \end{cases} \tag{1}$$

where R^{eq}_{t+1} and R^{bond}_{t+1} are the next-month returns for equities and bonds. We used the Vanguard Total World Stock ETF (VT) and the Vanguard Total Bond Market ETF (BND) as global stock and bond proxies. In the 48-month dataset, equities outperformed in 29 months (60.4%), establishing the majority-class baseline. The distribution reflects the broadly equity-positive period of 2020–2025, where stocks outperformed bonds every year, except in 2022.

IV. The CDSP Framework

Figure 1 provides an overview of the full pipeline, which consists of four stages: transcript preprocessing, chunk segmentation and context labeling, keyword extraction and taxonomy mapping, and feature construction.

A. Transcript Preprocessing

Raw transcripts contain substantial meeting logistics: greetings, audio-check exchanges, scheduling discussion, and speaker transitions. These segments are identified and removed by an LLM classifier prompted to distinguish administrative from substantive content. The cleaned transcript is then split into sentences, and portfolio decisions (if any) and filler words ("um," "you know," "uh") are removed. Apostrophes are preserved to maintain phrase integrity in the subsequent context window construction step.

B. Chunk Segmentation and Context Labeling

The cleaned transcript is then segmented into topical chunks, each representing a coherent discussion segment identified by a start indicator, that is, the first substantive sentence of the segment, extracted by the same LLM used for preprocessing. Typical chunks correspond to single topics such as "US equity valuation outlook" or "monetary policy trajectory". Each chunk is then assigned one or more context labels from the set: $\mathcal{C}$ ={equity, bonds, cross-asset, macroeconomic-outlook, meeting-logistic, other}.

Labeling is performed by an LLM prompted with label definitions and few-shot examples. The multi-label design reflects the comparative nature of cross-asset deliberation: a chunk discussing equity valuations relative to bond yields receives both *equity* and *cross-asset* labels.

Subsequently, for context-conditioning: a keyword occurrence in an *equity*-labeled chunk contributes to equity-context features, while the same keyword in a *macroeconomic-outlook* chunk contributes to macro-context features. These are treated as entirely distinct features. Chunks labeled *meeting-logistic* are excluded from all feature computation.

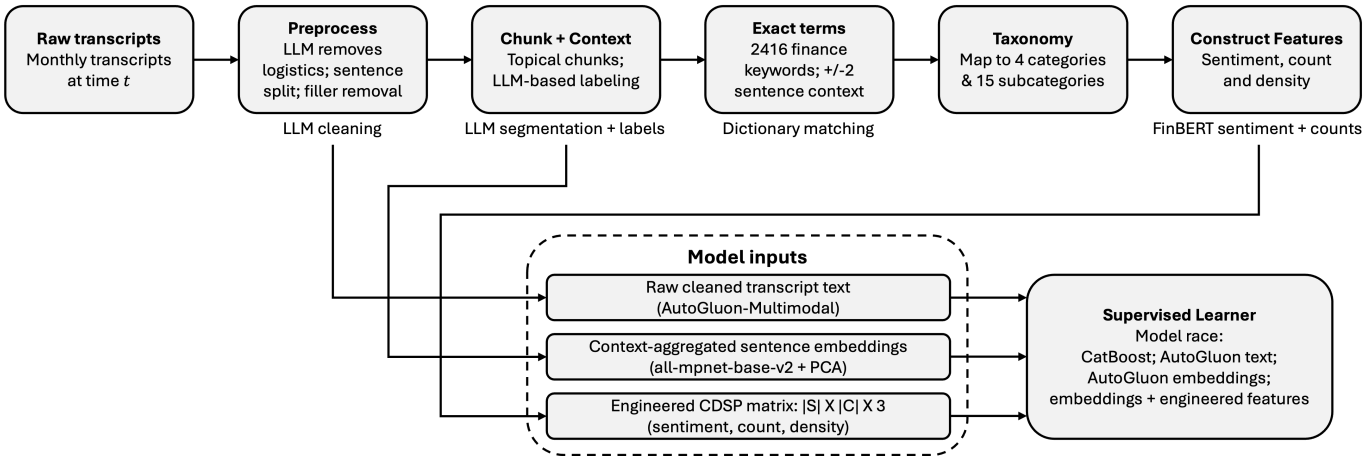

Fig. 1. CDSP pipeline for converting investment committee transcripts into context-conditioned features and predictive model inputs

TABLE I
KEYWORD TAXONOMY: CATEGORIES AND SUBCATEGORIES

Category	Subcategories
Portfolio Management	Asset Allocation Strategic Framework
Technical Analysis	Corporate Analysis Credit Analysis Comparative Analysis Quantitative Analysis
Economic Outlook	Economic Cycles Economic Indicators Forecasting Macroeconomic Monetary Policy Market Dynamics Geopolitical Factors
Risk Management	Macro Risk Risk Measurement

C. Keyword Extraction and Taxonomy Mapping

We maintain a vocabulary of 2,416 domain-specific financial keywords curated from investment committee language. The vocabulary covers the terminology of cross-asset portfolio management: asset allocation concepts ("diversification," "tactical positioning"), technical analysis terms ("yield curve," "spread tightening"), economic outlook language ("soft landing," "rate cycle," "geopolitical risk"), and risk management terminology ("duration," "hedging costs"). For each transcript, we record every occurrence of each vocabulary keyword, the chunk in which it appeared, and the chunk's context label(s). A context window of ± 2 sentences around each keyword is then constructed for sentiment scoring.

Keywords are mapped to a 4-category, 15-subcategory taxonomy (given in Table I), constructed iteratively through domain expert review. The taxonomy enables dimensionality reduction: rather than 2,416 individual keyword features, we aggregate to 15 subcategory-level signals.

D. Feature Construction

For each transcript month t, subcategory $s \in \mathcal{S}$ (15 subcategories), and asset-class context $c \in \mathcal{C}$, three feature types are constructed:

- **Sentiment score** $\sigma(t, s, c)$: the mean FinBERT [4] sentiment of all keyword context windows in subcategory s within context c in month t, normalized from the raw positive/negative/neutral probability outputs to a scalar in $[0, 1]$ (higher = more positive sentiment).
- **Mention count** $\phi(t, s, c)$: the total keyword occurrences in subcategory s within context c in month t.
- **Mention density** $\delta(t, s, c)$: $\phi(t, s, c)$ normalised by the total word count of chunks assigned to context c in month t, capturing relative discussion emphasis rather than absolute volume.

These features provide a month-level representation of each transcript that captures what topics were discussed, the contexts in which they appeared, and whether the surrounding language was positive or negative. For the classification experiments reported in this paper, we use only contemporaneous transcript-derived features from month t to predict the market direction in month $t + 1$.

E. Feature Modalities

Three modalities are used: (a) engineered features, (b) transcript text, and (c) sentence embeddings. The *engineered features* include the $\{\sigma, \phi, \delta\}$ variables described in the previous subsection. The *transcript text* includes full cleaned transcript text passed as a raw string to AutoGluon's multimodal predictor, which applies a transformer-based encoder for representational learning. The *sentence embedding* features include 768-dimensional chunk-level embeddings (computed using a sentence transformer [24], specifically `all-mpnet-base-v2`) aggregated to monthly vectors by context label. Principal component analysis (PCA) is then applied to reduce the overall dimensionality before training.

TABLE II
POINT-BISERIAL CORRELATIONS (r_{pb}) WITH NEXT-MONTH MARKET DIRECTION BY SUBCATEGORY AND CONTEXT ($N = 48$). BOLD: HIGHER $|r|$ BETWEEN ALL-CONTEXT AND EQUITY-CONTEXT. $^*p < 0.05$.

	Mention Count		Mention Density		Sentiment	
Subcategory	**All**	**Equity**	**All**	**Equity**	**All**	**Equity**
Asset Allocation	**+0.020**	-0.018	-0.067	**-0.070**	+0.219	**+0.294***
Strategic Framework	**+0.039**	-0.015	**-0.038**	+0.000	+0.070	**+0.080**
Corporate Analysis	-0.208	**-0.218**	-0.088	**-0.130**	-0.000	**+0.037**
Credit Analysis	**+0.156**	+0.037	-0.062	**+0.092**	**+0.050**	-0.004
Comparative Analysis	**+0.160**	+0.153	**+0.180**	+0.131	**+0.229**	+0.115
Quantitative Analysis	+0.027	**-0.037**	-0.118	**-0.153**	+0.185	**+0.191**
Economic Cycles	+0.019	**-0.141**	-0.013	**-0.197**	**+0.384***	+0.264
Econ. Indicators	+0.030	**+0.047**	+0.003	**-0.138**	**+0.174**	+0.164
Forecasting	-0.079	**-0.125**	**-0.367***	-0.180	**+0.282***	+0.186
Macroeconomic	+0.019	**-0.228**	-0.072	**-0.243**	**+0.241**	+0.131
Monetary Policy	+0.005	**+0.137**	**-0.059**	+0.009	+0.049	**+0.075**
Market Dynamics	**+0.101**	-0.086	**-0.175**	-0.149	**+0.224**	+0.062
Geopolitical	-0.072	**-0.284***	-0.165	**-0.252**	**+0.158**	+0.021
Risk Measurement	+0.046	**-0.067**	-0.090	**-0.150**	**+0.160**	+0.082
Macro Risk	**-0.156**	-0.108	-0.116	**-0.130**	+0.004	**-0.023**

V. FEATURE DESCRIPTION

A. Mention Frequency vs. Sentiment as Predictors

Table II reports point-biserial correlations (r_{pb}) of mention count, mention density, and sentiment score features with the binary market direction target ($y_t = 1$: equity outperforms), for each subcategory. These are shown separately for "All" (all chunks, regardless of context labels) and "Equity" (restricted just to chunks with the equity context label). Bold values indicate the higher $|r_{pb}|$ between the two contexts. The large amount of text in these transcripts enables the CDSP approach. However, given the fact that the committee meets less than monthly, this limits the sample size (N = 648) and the statistical significance of the correlations.

The sentiment feature displays stronger correlations with market direction than mention frequency (count and density) across most subcategories. Among sentiment features, Economic Cycles, Asset Allocation, and Forecasting show the largest positive correlations, indicating that more positive deliberation sentiment in these categories precedes equity outperformance months. On the count channel, the strongest signals are negative: Geopolitical, Macroeconomic, and Corporate Analysis count features are associated with bond outperformance months.

B. Context-Conditioning Effect

Does context-conditioning improve the correlation with market direction? We see that this is the case more in the mention frequency features (count and density), but it seems to degrade correlation in the case of the sentiment feature. Therefore, context-conditioning appears to be helpful for some features, but not all.

C. Topic Attention Over Time

Figure 2 overlays the two strongest sentiment features, that is, Economic Cycles and Asset Allocation, against the monthly market direction outcome to illustrate how the top signals track the equity/bond regime over time.

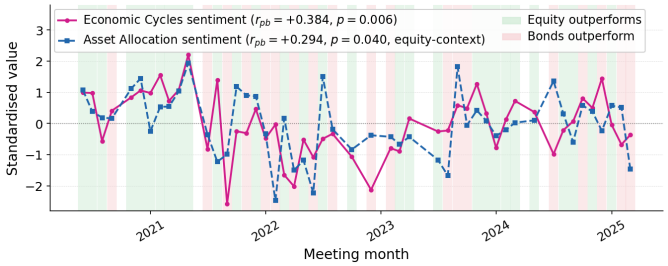

Fig. 2. Standardised Economic Cycles sentiment ($r_{pb} = +0.384$, $p = 0.006$) and Asset Allocation sentiment ($r_{pb} = +0.294$, $p = 0.040$, equity-context) overlaid on next-month market direction outcomes (Jun 2020–Feb 2025). Green background: equity outperforms; red: bonds outperform.

The two features confirm the pattern from Table II: more positive deliberation sentiment in Economic Cycles and Asset Allocation subcategories tends to co-occur with subsequent equity outperformance. Neither feature tracks the outcome perfectly: both signals are noisy and the correlations are moderate. But the directional tendency is visible across the sample, including during the 2022 bond-outperformance episodes where both sentiment features are notably suppressed.

VI. PREDICTIVE MODELS

A. Experimental Setup

The model setup includes: (1) *CatBoost* [27] gradient boosted classifier with L_2 regularisation, trained on engineered semantic features only. (2) *AutoGluon* trained on raw transcript text only, which internally applies a transformer encoder to produce document representations. (3) AutoGluon trained on sentence embeddings alone. (4) AutoGluon trained on sentence embeddings combined with engineered features.

All models are evaluated under 5×5 repeated stratified k-fold cross-validation. Each test has five repetitions with different random seeds to stabilise variance estimates on the small ($N = 48$) dataset, producing a distribution of fold-level scores rather than a single estimate.

Two baselines are used: (1) *Majority-class*: always predict equity outperforms (60.4% accuracy, 0.449 F1). (2) *Lag-1 direction*: predict the same market outcome as the previous month (52.1% accuracy, 0.494 F1). The lag-1 baseline underperforms majority-class accuracy because monthly outcomes have low serial correlation, that is, last month's winner is not a reliable predictor for this month.

B. Results

Table III reports accuracy and F1 for all models and baselines, and Figure 3 visualizes the comparison. The majority-class baseline strategy achieves 60.4% accuracy but only 0.449 F1, reflecting near-zero recall on the minority (bonds-outperform) class. All four predictive models surpass this on F1, i.e., the transcript signal improves minority-class identification over baseline. Furthermore, the lag-1 baseline is weak, achieving only 52.1% accuracy, which is even below

TABLE III
MARKET DIRECTION CLASSIFICATION RESULTS
(5×5 REPEATED STRATIFIED K-FOLD, $N = 48$)

Type	Model / Features	Acc.	F1
Baseline	Majority class	0.604	0.449
	Lag-1 direction	0.521	0.494
Model	CatBoost (Eng. features)	0.620	0.620
	AutoGluon (Transcripts)	0.710	0.690
	AutoGluon (Embeddings)	0.630	0.610
	AutoGluon (Embed.+Feat.)	**0.730**	**0.730**

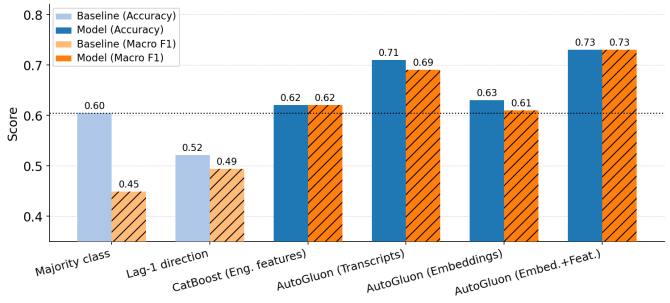

Fig. 3. Accuracy (solid) and F1 (hatched) for all models and baselines. The dotted line marks the 60.4% majority-class baseline. AutoGluon with combined embeddings and engineered features achieves a 12.6 percentage point improvement over the majority baseline.

the majority-class baseline. This suggests that the predictive models, outperforming both baselines, are not just reflecting forward persistence in the target. Few other takeaways include:

- AutoGluon on transcripts alone achieves 71% accuracy and 0.69 F1, without any feature engineering. This result suggests that the predictive signal is a property of the deliberation text, as opposed to an artifact of our specific taxonomy or feature construction choices.
- Combining sentence embeddings with engineered features lifts accuracy from 63% to 73% and F1 from 0.61 to 0.73. The engineered features capture structured, interpretable patterns (which subcategories are discussed, in which context, at what volume) that are not fully encoded in the dense embedding representation. The two modalities extract complementary information from the same source.
- CatBoost with engineered features only is the most interpretable model configuration. Without raw text or embeddings, the semantic features produce 62% accuracy and 0.62 F1, thus improvement over the baselines is traceable to named subcategories.

C. Feature Importance

Figure 4 shows feature importances from the CatBoost model trained on engineered features. The most important feature is Corporate Analysis density, followed by equity-context Risk Management density and equity-context Asset Allocation density. The dominance of density features over raw count features is consistent with the correlation analysis: relative discussion emphasis is more informative than absolute keyword

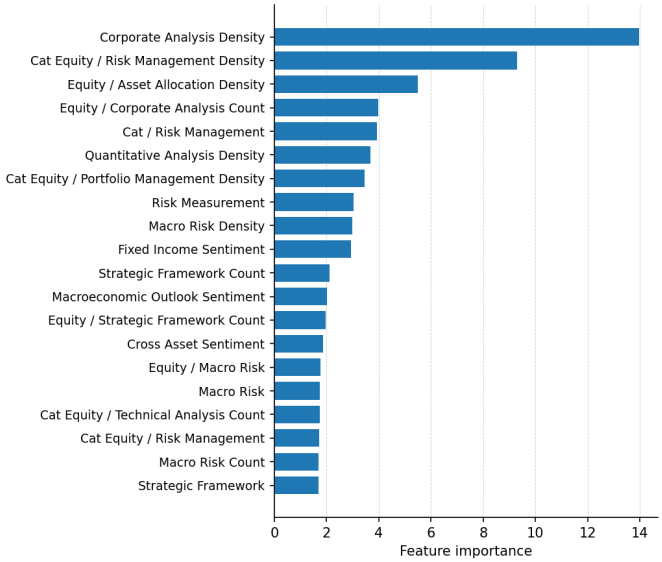

Fig. 4. Top 20 CatBoost feature importances for market direction classification (engineered features only).

volume for some categories (see Table II). The prominence of Risk Management and Asset Allocation features in equity-context specifically confirms that the same concept discussed during an equity-framed segment is more predictive than the same concept discussed in a general context, supporting the context-conditioning design of CDSP.

VII. DISCUSSION

A. Two Complementary Signal Channels

The empirical results reveal two distinct and complementary signal mechanisms in the transcripts.

The first is *structured deliberation signal*: what topics are discussed, how much attention they receive, and the sentiment with which they are discussed are all predictive of subsequent market direction. This channel is captured by the engineered features and is interpretable at the subcategory level. More positive Economic Cycles and Asset Allocation sentiment precedes equity outperformance; elevated Geopolitical and Macroeconomic discussion volume precedes bond outperformance. The sentiment channel dominates both count and density for several subcategories, while context-conditioning (equity-framed versus general) substantially strengthens the geopolitical and macroeconomic count/density signals.

The second is *latent deliberation content*: the full transcript text, passed to a neural model without any feature engineering, achieves 71% accuracy. This suggests the signal is richer than what the 15-subcategory taxonomy encodes: nuance in phrasing, speaker emphasis, the balance of hedged versus assertive language, and cross-sentence reasoning all contribute, even if they are not individually labeled. However, in regulated financial settings, predictive performance alone is insufficient: decisions must be explainable to risk managers, compliance officers, and investment committees. Engineered features satisfy this requirement directly: a prediction driven by elevated

equity-context Geopolitical count or positive Economic Cycles sentiment can be traced back to specific discussion segments in the original transcript, whereas embedding-based models offer no easy attribution path.

The fact that combining both channels (73%) improves on either alone (71% and 63%) suggests they are complementary rather than redundant. For practitioners, this suggests a two-layer architecture: an interpretable structured-feature model for explainability and a neural model for performance, combined at inference.

B. Practical Implications

The fact that the dual-channel method had a 73% accuracy for our task of selecting the correct monthly equity-versus-bonds choice whereas the method of always selecting equities had only an accuracy of 60.4% may be practically meaningful, but we also note this result is statistically weak. More specifically, applying McNemar's test to the same data yields a p-value of 0.263, largely due to the fact that we only have 48 data points. In a portfolio management context, better directional accuracy, applied to a binary risk-on/risk-off decision, may be useful despite market efficiency making prediction difficult, particularly in volatile regimes where the majority-class heuristic (i.e., always choosing equities) performs poorly. The finding that the signal persists across 2022 (a bond-outperformance year) and 2023–2024 (equity-outperformance years) suggests the model is capturing regime-varying features rather than a single-regime shortcut.

More broadly, the result supports the "soft information" hypothesis [21], [22]: qualitative deliberation encodes forward-looking information that is not already captured in the quantitative inputs available at the time of the meeting.

C. Generalizability

The CDSP pipeline is domain-agnostic. The four design elements (that is, chunk segmentation, LLM context labeling, taxonomy-structured extraction, and feature construction) transfer to any setting where experts deliberate over allocation or directional decisions in a transcribed, topically structured way. The only necessary contextual customization is creating the keyword taxonomy, which can be curated from the available dataset using any standard LLM model, and reviewed by domain experts for contextual relevance. While there are organic extensions of this framework to other domains in finance such as analyst calls deliberating stock related decisions, real-estate analysts deliberating construction pipelines, etc., there are possible extensions to non-finance domains as well, for example, management meetings.

D. Limitations

Dataset size: $N = 48$ observations is the primary constraint on all results. The 5×5 repeated stratified CV mitigates fold-level variance, but each fold contains 38–40 training examples, with the risk of overfitting. The dominance of linear and simple tree models over deep ensembles in our experiments (AutoGluon's best models were typically linear) is consistent with this constraint.

Transcript quality: Automatic speech recognition introduces errors that reduce keyword recall, particularly for technical terms, acronyms (e.g., "VIX," "GDP," "EPS"), and proper nouns. Estimated recall impact is 10–15% of keywords in a representative set.

Temporal structure: The 5×5 stratified CV does not preserve temporal ordering. A strict walk-forward evaluation, training on months 1–40 and evaluating on months 41–48, would be a more conservative out-of-time test, though with only 8 holdout points. Future work with a larger dataset should prioritize temporal evaluation.

VIII. Conclusion

This paper introduced CDSP, a context-conditional deliberation signal pipeline for transforming investment committee discussions into structured, predictive signals. The central idea is that expert discussion should not be treated as a flat text document. The same financial concept can carry different meaning depending on whether it appears in an equity discussion, a bond discussion, a macroeconomic overview, etc. CDSP preserves this structure by segmenting transcripts into topical chunks, assigning context labels, mapping keywords to a financial taxonomy, and constructing sentiment and attention-based features within each context.

Applied to 48 investment committee transcripts, the framework produces three main findings. *First*, sentiment polarity has higher correlations to next-month market direction than mention frequency. *Second*, the transcript language contains some information beyond the engineered taxonomy. Raw transcript models perform well on their own, and the best result is achieved by combining embeddings with CDSP features, reaching an accuracy of 73% against a 60.4% majority-class (all equity) baseline over our 48 investment committee transcripts.

The broader contribution is methodological. CDSP offers a way to convert qualitative expert reasoning into features that are both empirically testable and partially interpretable. This is important in financial settings where performance alone is insufficient and predictions must be traceable to recognizable topics, contexts, and discussion patterns.

Overall, the evidence suggests that investment committee discussions may encode forward-looking information about market regimes. NLP provides a practical way to recover that information, linking the qualitative structure of expert deliberation to supervised modeling. The same approach could be extended to other settings, such as buy-side/sell-side decisions, real-estate planning, medical discussions, and other expert forums where spoken reasoning is recorded and outcomes can be quantified.

Acknowledgements

[Omitted for review.]

REFERENCES

[1] P. C. Tetlock, "Giving content to investor sentiment: The role of media in the stock market," The Journal of Finance, vol. 62, no. 3, pp. 1139–1168, 2007.

[2] T. Loughran and B. McDonald, "When is a liability not a liability? textual analysis, dictionaries, and 10-Ks," The Journal of Finance, vol. 66, no. 1, pp. 35–65, 2011.

[3] M. Druz, I. Petzev, A. F. Wagner, and R. J. Zeckhauser, "When Managers Change Their Tone, Analysts and Investors Change Their Tune," Financial Analysts Journal, vol. 76, no. 2, pp. 47–69, 2020. [Online]. Available: https://doi.org/10.1080/0015198X.2019.1707592

[4] D. Araci, "FinBERT: Financial sentiment analysis with pre-trained language models," 2019. [Online]. Available: https://arxiv.org/abs/1908.10063

[5] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann, "BloombergGPT: A large language model for finance," 2023. [Online]. Available: https://arxiv.org/abs/2303.17564

[6] A. López-Lira and Y. Tang, "Can ChatGPT forecast stock price movements? return predictability and large language models," 2023. [Online]. Available: https://arxiv.org/abs/2304.07619

[7] S. Hansen, M. McMahon, and A. Prat, "Transparency and deliberation within the FOMC: A computational linguistics approach," The Quarterly Journal of Economics, vol. 133, no. 2, pp. 801–870, 2018.

[8] E. Osowska and P. Wójcik, "Predicting the reaction of financial markets to Federal Open Market Committee post-meeting statements," Digital Finance, vol. 6, no. 1, pp. 145–175, 2024.

[9] K. Du, Y. Zhao, R. Mao, F. Z. Xing, and E. Cambria, "Natural language processing in finance: A survey," Information Fusion, vol. 115, p. 102755, 2025.

[10] S. Kogan, D. Levin, B. R. Routledge, J. S. Sagi, and N. A. Smith, "Predicting risk from financial reports with regression," in Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Boulder, Colorado: Association for Computational Linguistics, June 2009, pp. 272–280. [Online]. Available: https://aclanthology.org/N09-1031/

[11] Z. T. Ke, B. T. Kelly, and D. Xiu, "Predicting returns with text data," National Bureau of Economic Research, NBER Working Paper 26186, August 2019. [Online]. Available: https://www.nber.org/papers/w26186

[12] S. R. Das and M. Y. Chen, "Yahoo! for Amazon: Sentiment extraction from small talk on the web," Management Science, vol. 53, no. 9, pp. 1375–1388, 2007.

[13] J. Bollen, H. Mao, and X.-J. Zeng, "Twitter mood predicts the stock market," Journal of Computational Science, vol. 2, no. 1, pp. 1–8, 2011.

[14] D. F. Larcker and A. A. Zakolyukina, "Detecting deceptive discussions in conference calls," Journal of Accounting Research, vol. 50, no. 2, pp. 495–540, 2012.

[15] W. Y. Wang and Z. Hua, "A semiparametric gaussian copula regression model for predicting financial risks from earnings calls," in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Baltimore, Maryland: Association for Computational Linguistics, June 2014, pp. 1155–1165. [Online]. Available: https://aclanthology.org/P14-1109/

[16] F. Z. Xing, E. Cambria, and R. E. Welsch, "Natural language based financial forecasting: A survey," Artificial Intelligence Review, vol. 50, no. 1, pp. 49–73, 2018.

[17] L. Malandri, F. Z. Xing, C. Orsenigo, C. Vercellis, and E. Cambria, "Public mood-driven asset allocation: The importance of financial sentiment in portfolio management," Cognitive Computation, vol. 10, no. 6, pp. 1167–1176, 2018.

[18] F. Z. Xing, E. Cambria, and R. E. Welsch, "Growing semantic vines for robust asset allocation," Knowledge-Based Systems, vol. 165, pp. 297–305, 2019.

[19] X. Du and K. Tanaka-Ishii, "Stock embeddings acquired from news articles and price history, and an application to portfolio optimization," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, July 2020, pp. 3353–3363. [Online]. Available: https://aclanthology.org/2020.acl-main.307/

[20] M.-C. Hung, P.-H. Hsia, X.-J. Kuang, and S.-K. Lin, "Intelligent portfolio construction via news sentiment analysis," International Review of Economics & Finance, vol. 89, pp. 605–617, 2024.

[21] J. M. Liberti and M. A. Petersen, "Information: Hard and soft," The Review of Corporate Finance Studies, vol. 8, no. 1, pp. 1–41, 2019.

[22] J. Wang, "Screening soft information: Evidence from loan officers," RAND Journal of Economics, vol. 51, no. 4, pp. 1287–1322, 2020.

[23] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," Journal of Machine Learning Research, vol. 3, pp. 993–1022, 2003. [Online]. Available: https://jmlr.org/papers/v3/blei03a.html

[24] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, November 2019, pp. 3982–3992. [Online]. Available: https://aclanthology.org/D19-1410/

[25] N. Erickson, J. Mueller, A. Shirkov, H. Zhang, P. Larroy, M. Li, and A. Smola, "AutoGluon-Tabular: Robust and accurate AutoML for structured data," 2020. [Online]. Available: https://arxiv.org/abs/2003.06505

[26] Z. Tang, H. Fang, S. Zhou, T. Yang, Z. Zhong, C. Hu, K. Kirchhoff, and G. Karypis, "AutoGluon-Multimodal (AutoMM): Supercharging multimodal AutoML with foundation models," in Proceedings of the Third International Conference on Automated Machine Learning, ser. Proceedings of Machine Learning Research, K. Eggensperger, R. Garnett, J. Vanschoren, M. Lindauer, and J. R. Gardner, Eds., vol. 256. PMLR, 09–12 Sep 2024, pp. 15/1–35. [Online]. Available: https://proceedings.mlr.press/v256/tang24a.html

[27] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in Advances in Neural Information Processing Systems, vol. 31, 2018.