CONVERSE-CAUSE: Causal Analysis and Interactive Reasoning over Conversational Data Using Retrieval-Augmented Evidence

Shreeya Rani Das,Bhabesh Behera,Ankit Kumar Das,Pritam Prayash Behera,

Team Sages/ITER-SOA

Manuscript received On Feb Month 6th Day, 2026. This work was developed as part of a data science research project on conversational causal analysis.

Abstract: Large-scale conversational systems generate extensive multi-turn dialogue data associated with operational outcomes such as fraud investigations, delivery issues, account access failures, and service escalations. While these outcomes are recorded, the conversational factors leading to them are often not explicitly identified. This work presents an outcome-agnostic framework that combines interpretable machine learning with retrieval-augmented evidence extraction to identify causal conversational patterns and supporting dialogue turns. The system further enables interactive multi-turn analysis with deterministic context memory, ensuring explanations remain consistent and grounded in the original transcripts. The proposed approach improves outcome prediction accuracy while providing transparent, evidence-based reasoning over conversational data.

Index Terms: Conversational analytics, causal reasoning, retrieval-augmented generation, explainable AI, dialogue systems.

1. Introduction

Modern customer support and service platforms generate large volumes of multi-turn conversational data across domains such as banking, telecommunications, healthcare, and e-commerce. These conversations are often associated with operational outcomes including fraud investigations, delivery issues, account access failures, complaints, and service escalations. While organizations routinely record these outcomes, the conversational factors that contribute to them remain largely unobserved and difficult to analyze.

Traditional conversational analytics primarily focus on intent classification, sentiment analysis, or outcome prediction. Although these methods can identify what outcome occurred, they do not explain *why* the outcome happened or which specific dialogue behaviors contributed to it. This lack of interpretability limits their usefulness for operational improvement, quality monitoring, and decision support.

Recent advances in machine learning and natural language processing have improved the ability to extract semantic information from conversations. However, many high-performing models operate as black boxes and often generate explanations that are not grounded in the original dialogue data. For operational environments, it is essential that analytical insights are both interpretable and traceable to concrete conversational evidence.

In this work, we present **CONVERSE-CAUSE**, an outcome-agnostic framework for causal analysis and interactive reasoning over conversational transcripts. The proposed system combines

interpretable machine learning with retrieval-based evidence extraction to identify conversational patterns associated with different operational outcomes. A Retrieval-Augmented Generation (RAG) pipeline is used to retrieve relevant dialogue turns that support causal explanations, ensuring faithfulness to the underlying data.

In addition to single-query analysis, the system supports multi-turn analytical interaction through a deterministic context memory mechanism. This enables users to ask follow-up questions such as identifying the most influential dialogue turn, comparing contributing factors, or examining early warning signals, while maintaining consistency across interactions.

The main contributions of this work are as follows:

- An outcome-agnostic conversational analytics framework that supports multiple operational event types.
- A hybrid modeling approach combining semantic text features and interpretable behavioral features for accurate and explainable outcome prediction.
- A retrieval-based evidence mechanism that links causal explanations to specific dialogue turns.
- A context-aware interaction module that enables consistent multi-turn analytical reasoning.

The proposed approach improves both predictive performance and interpretability, enabling organizations to move from simple event detection to actionable causal understanding of conversational behavior.

2. Dataset and Problem Formulation

Large-scale conversational systems generate multi-turn dialogue data across multiple operational domains. Each conversation represents an interaction between a customer and an agent and is associated with a final operational outcome. In this work, the outcome is represented by the *intent* field, which is treated as a multi-class target variable. The objective is twofold: (1) predict the operational outcome using conversational signals, and (2) identify causal conversational factors with evidence-based explanations.

The dataset used in this study contains structured conversational records with the following attributes: transcript identifier, timestamp of interaction, domain, reason for call, and ordered dialogue turns with speaker labels. The data spans multiple domains including banking, customer support, and service operations, allowing evaluation of outcome-agnostic modeling approaches.

The problem is formulated as a supervised multi-class classification task combined with a retrieval-based explanation task. Given a conversation, the system predicts the most likely operational outcome and retrieves dialogue segments that provide evidence for the prediction. This formulation enables both predictive analytics and interpretable causal reasoning.

The framework is designed to operate under realistic operational constraints, including noisy text, variable conversation length, and domain diversity. The proposed approach integrates statistical learning with semantic retrieval to provide both accuracy and interpretability.

3. Data Cleaning and Exploratory Analysis

Data preprocessing plays a critical role in conversational analytics due to inconsistencies in real-world dialogue data. Initial cleaning steps included removal of duplicate records, handling of missing values, and normalization of textual content. All conversation text was converted to lowercase and non-informative characters were removed to ensure consistent tokenization.

Multi-turn conversations were flattened into structured sequences while preserving speaker order. Turn-level segmentation was also maintained for downstream retrieval tasks. Timestamp fields were converted into temporal features such as hour-of-day to capture operational patterns related to workload and service availability.

Exploratory Data Analysis (EDA) was conducted to understand conversational behavior across outcomes. Distribution analysis revealed class imbalance across operational intents, which was

addressed using stratified sampling during model training. Conversation length analysis showed that adverse outcomes were often associated with longer interactions.

Speaker interaction patterns were analyzed using customer-to-agent turn ratios. Results indicated that conversations with high customer dominance often corresponded to unresolved issues or dissatisfaction. Sentiment analysis further revealed that negative polarity accumulation across turns correlated with complaint-oriented outcomes.

These insights guided the design of outcome-agnostic behavioral features that capture conversational dynamics rather than domain-specific patterns.

4. Feature Engineering and Preprocessing

Feature engineering was designed to capture both structural interaction signals and semantic conversational context. Outcome-agnostic behavioral features include total number of turns, customer turn ratio, average customer sentiment polarity, frequency of issue-related keywords, and time-of-interaction features.

Customer sentiment polarity was computed using lexicon-based analysis over customer turns only, ensuring that the feature reflects user experience rather than agent responses. Outcome signal frequency was calculated using a generic keyword set including terms such as *issue*, *problem*, *error*, *delay*, and *refund*.

To capture semantic information, conversation text was transformed using Term Frequency–Inverse Document Frequency (TF-IDF) vectorization with unigram and bigram features. The TF-IDF representation provides interpretable text features while maintaining computational efficiency for large datasets.

Numerical features were standardized using z-score normalization, and extreme values were handled using winsorization between the 5th and 95th percentiles. The final feature space was constructed by combining sparse TF-IDF vectors with scaled behavioral features.

The target variable was encoded using label encoding, and the dataset was split into training and testing sets using stratified sampling (80/20) to preserve class distribution.

5. Outcome Prediction Models

Multiple machine learning models were evaluated to identify an optimal balance between performance and interpretability. Logistic Regression with a one-vs-rest strategy was used as a baseline due to its transparency and ability to provide feature coefficients for causal interpretation.

Tree-based ensemble methods including Random Forest and Gradient Boosting were also evaluated to capture non-linear relationships between conversational features and operational outcomes. These models are robust to noise and capable of handling mixed feature types.

Model performance was evaluated using accuracy, precision, recall, and F1-score. Cross-validation confirmed that hybrid feature representations combining semantic and behavioral signals significantly improved predictive performance compared to either feature type alone.

Feature importance analysis revealed that semantic features contributed to outcome discrimination, while behavioral features provided interpretable signals related to conversational dynamics. This separation supports the dual objective of accuracy and explainability.

6. Embedding-Based Semantic Modeling

To support retrieval-based reasoning, turn-level semantic embeddings were generated using Sentence-BERT (all-MiniLM-L6-v2). This transformer-based model produces dense vector representations that preserve semantic similarity between dialogue segments.

Each dialogue turn was encoded independently and stored along with metadata including transcript identifier, speaker role, and turn position. The embeddings enable fine-grained semantic search across large conversational corpora.

Vector similarity search was implemented using Facebook AI Similarity Search (FAISS), which

provides efficient indexing and retrieval for high-dimensional vectors. The FAISS index enables real-time retrieval of relevant dialogue segments based on user queries or model explanations.

Embedding-based retrieval complements statistical modeling by enabling context-aware evidence extraction that is not limited to predefined keywords or features.

7. Retrieval-Augmented Causal Reasoning

A Retrieval-Augmented Generation (RAG) framework was implemented to ensure that explanations remain grounded in the original conversational data. Evidence retrieval is guided by two criteria: causal importance and semantic relevance.

Causal importance is derived from model feature weights and interaction signals, identifying conversational regions most likely to influence the predicted outcome. Semantic relevance is computed using embedding similarity between user queries and dialogue turns.

The retrieved evidence is used to construct structured explanations that include the predicted outcome, primary contributing factors, and supporting dialogue excerpts. This approach minimizes hallucination and ensures traceability.

The framework also supports deterministic multi-turn interaction through a context memory that stores active transcript information, identified causal factors, and previously retrieved evidence. Follow-up queries reuse this context to maintain consistency across analytical sessions.

8. Results and Discussion

The proposed framework was evaluated on a held-out test set. Outcome prediction results are summarized as follows:

- Logistic Regression Accuracy: 92.4%
- Random Forest Accuracy: 93.1%
- Gradient Boosting Accuracy: 94.6%

The results demonstrate that hybrid semantic-behavioral feature representations significantly improve classification performance. Precision and recall values remained consistent across major outcome classes, indicating robust generalization.

Embedding-based retrieval achieved high semantic relevance, with retrieved turns consistently corresponding to conversational segments preceding the operational outcome. Qualitative analysis confirmed that explanations aligned with observable conversational patterns such as repeated issue descriptions or escalation language.

The interactive reasoning component successfully supported follow-up queries without recomputation, maintaining deterministic context and improving user interpretability.

9. Conclusion

This paper presented CONVERSE-CAUSE, a scalable framework for causal analysis and interactive reasoning over conversational data. By integrating interpretable machine learning, semantic embeddings, and retrieval-based evidence, the system achieves high predictive accuracy while maintaining transparency and traceability.

The proposed approach enables organizations to move beyond outcome detection toward actionable causal insights. Future work includes deployment in real-time monitoring systems and the integration of counterfactual conversational analysis.

References

References

[1] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[3] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” in *Proc. EMNLP*, 2019, pp. 3982–3992.

[4] J. Johnson, M. Douze, and H. Jégou, “Billion-scale similarity search with GPUs,” *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2021.

[5] P. Lewis et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” in *Proc. NeurIPS*, 2020.

[6] R. Guidotti et al., “A Survey of Methods for Explaining Black Box Models,” *ACM Computing Surveys*, vol. 51, no. 5, pp. 1–42, 2018.

[7] Hugging Face, “Transformers: State-of-the-art Machine Learning for PyTorch, TensorFlow, and JAX,” Available: *https://huggingface.co*. Accessed: Jan. 2026.

[8] Kaggle Inc., “Kaggle: Your Machine Learning and Data Science Community,” Available: *https://www.kaggle.com*. Accessed: Jan. 2026.

[9] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[10] Streamlit Inc., “Streamlit: The fastest way to build data apps in Python,” Available: *https://streamlit.io*. Accessed: Jan. 2026.