Natural Language Processing for Healthcare M&A Communication Analysis: Predicting Deal Success Through Linguistic Pattern Recognition

Maksim Kocheshkov
Master's Program in Data Science
University of Milan
[maksim.kocheshkov@studenti.unimi.it]

July 23, 2025

Abstract

This study analyzes healthcare merger and acquisition (M&A) announcement communications using natural language processing techniques. 193 healthcare M&A transactions from 2021-2025 were analyzed, applying domain-specific word embeddings, concept activation vectors (CAVs), temporal analysis, and explainable AI methods to understand linguistic patterns in deal communications.

Key findings reveal that strategic value communication (SHAP importance: 0.041) and financial strength messaging (0.0409) are the strongest predictors of communication patterns, with integration focus showing significant individual prediction impact. The temporal analysis demonstrates evolution in communication strategies across market cycles, with risk factor emphasis varying substantially between the post-COVID boom period (2021) and subsequent market corrections (2022-2024). Explainability analysis using SHAP provides actionable insights for M&A professionals. The study contributes methodological innovation in domain-specific NLP and practical business intelligence for healthcare M&A strategy.

Keywords: Natural Language Processing, Mergers & Acquisitions, Healthcare, Text Mining, SHAP Analysis, Word Embeddings

Contents

1 Introduction

Healthcare M&A activity has surged in recent years, yet the linguistic patterns in deal announcements remain unexplored, representing a gap at the intersection of NLP and finance. Traditional M&A research focuses on quantitative metrics, overlooking communication strategies that may contain predictive signals.

Healthcare transactions present unique challenges, navigating complex regulatory environments and diverse stakeholder groups. This study addresses: (1) What linguistic patterns characterize healthcare M&A communications? (2) How do communication strategies evolve over time? (3) Which business concepts can be extracted and analyzed?

I analyze 193 healthcare M&A transactions (2021-2025) using domain-specific Word2Vec embeddings, Concept Activation Vectors (CAVs), temporal analysis, and SHAP explainability. The contribution demonstrates how domain-specific NLP techniques capture nuanced business concepts beyond traditional sentiment analysis.

2 Research Question and Methodology

2.1 Methodological Framework

The methodology integrates multiple NLP techniques designed for business domain analysis.

2.1.1 Phase 1: Foundation Analysis and Business Feature Extraction

I extract 17 M&A-specific business features using BERT-compatible tokenization with intelligent chunking (50-token overlap) to handle documents exceeding 512-token limits. Features span: (1) Strategic rationale indicators (market expansion, product diversification, technology acquisition), (2) Financial communication patterns (confidence ratios), (3) Risk factors (regulatory mentions, complexity), (4) Healthcare-specific elements (clinical integration, regulatory pathways), (5) Communication tone indicators.

2.1.2 Phase 2: Domain-Specific Embedding Development

Custom Word2Vec embeddings trained exclusively on the M&A corpus use: 200 dimensions, 8-word context window, minimum count 5, 100 epochs, 15 negative samples. This configuration optimizes for specialized business vocabulary and relationship patterns. Semantic validation examines business relationships like acquisition-transaction and synergies-cost savings.

2.1.3 Phase 3: Concept Activation Vector (CAV) Analysis

Following Kim et al., I implement CAVs, but considering industry specificity. 6 business concepts were extracted: Financial Strength, Strategic Value, Risk Factors, Integration Focus, Stakeholder Focus, and Deal Complexity.

2.1.4 Phase 4: Temporal Analysis

I analyze communication evolution across five market periods: 2021 Post-COVID Boom, 2022 Interest Rate Correction, 2023 Market Stabilization, 2024 Strategic Consolidation, and 2025 Future Focus. This examines how external market forces influence communication strategies.

2.1.5 Phase 5: Explainable AI

SHAP analysis using Random Forest classifiers and TreeExplainer calculates feature importance, ensuring interpretability for business practitioners.

2.2 Problem Definition

The analysis extracts linguistic patterns from M&A announcements. Given corpus $D = \{d_1, d_2, ..., d_n\}$ of n documents with feature vectors $\mathbf{x}_i \in \mathbb{R}^m$, I learn representations $f : \mathbb{R}^m \to \mathbb{R}^k$ mapping features to interpretable business concept scores.

3 Experimental Results

3.1 Dataset Overview

193 healthcare M&A transactions (2021-2025) from Refinitiv include: acquisitions (94, 48.7%), mergers (21, 10.9%), asset acquisitions (16, 8.3%), licensing (12, 6.2%), reverse mergers (10, 5.2%). Geographic distribution: US 72.2%, Europe 14.4%, India 5.6%, Japan 3.7%.

3.2 Business Feature Results

Strategic patterns: product diversification (66 deals, 34.2%), market expansion (60 deals, 31.1%), technology acquisition (50 deals, 25.9%). Financial confidence ratio averages 0.53 with acquisitions showing higher confidence (0.58) than mergers (0.51). Regulatory mentions average 0.6 per document. Clinical integration appears in 78% of clinical asset transactions.

3.3 Domain-Specific Word Embedding Analysis

Training custom Word2Vec embeddings on the M&A corpus yielded significant improvements in semantic representation quality. The final vocabulary contains 9,352 unique terms after filtering, representing a 52% retention rate from the initial 19,745 terms.

Semantic Relationship Validation: Analysis of learned embeddings demonstrates successful capture of M&A-specific semantic relationships:

- **Transaction Structure:** "acquisition" shows highest similarity to "transaction" (0.67), "proposed" (0.61), and "deal" (0.58)
- **Financial Metrics:** "accretive" demonstrates strong associations with "eps" (0.71), "accretion" (0.68), and "earnings" (0.63)
- **Strategic Value:** "synergies" correlates highly with "synergy" (0.74), "cost" (0.59), and "incremental" (0.56)

- **Valuation Language:** "premium" shows appropriate relationships with "price" (0.62), "unaffected" (0.58), and "vwap" (0.55)

3.4 CAV Analysis Results

Six business concepts extracted with 95-100% classifier accuracy. Concept emphasis patterns: Risk Factors dominant (0.245), Integration Focus second (0.169), Strategic Value negative (-0.089), Financial Strength negative (-0.112), Stakeholder Focus lowest (-0.151), Deal Complexity most negative (-0.218). Strong negative correlation between strategic messaging and risk discussion ($r = -0.42$) indicates communication trade-offs.

3.5 Temporal Analysis Results

Communication evolution across market periods: **2021** (19 deals): High financial confidence (0.61), technology focus peaked. **2022** (58 deals): Increased risk emphasis (0.257), reduced confidence (0.520), product diversification surge. **2023** (47 deals): Risk communication decreased (0.234), strategic messaging improved. **2024** (74 deals): Lowest confidence (0.420), peak integration focus (0.180). **2025** (18 deals): Renewed strategic focus (-0.066), reduced risk emphasis (0.249).

3.6 SHAP Explainability Analysis Results

SHAP analysis provides detailed insights into which features most strongly characterize communication patterns.

Feature Importance Rankings:

1. **Strategic Value (SHAP importance: 0.0410):** Highest global importance, indicating that strategic communication quality is the strongest characterizing feature
2. **Financial Strength (SHAP importance: 0.0409):** Nearly equal importance to strategic value
3. **Integration Focus (SHAP importance: 0.0402):** Third-highest importance, validating the significance of integration planning discussion
4. **Deal Complexity (SHAP importance: 0.0394):** Moderate importance with negative correlation patterns
5. **Risk Factors (SHAP importance: 0.0368):** Lowest importance among major features

4 Concluding Remarks

This research demonstrates that healthcare M&A communications contain identifiable linguistic patterns extractable via domain-specific NLP techniques. Strategic value communication and financial strength messaging emerge as primary characteristics, while risk factor discussion dominates but shows lower predictive significance.

Methodological Contributions: (1) Domain-specific embeddings outperform generic models for business applications, (2) CAVs successfully bridge linguistic features and business concepts, (3) Temporal integration captures market context effects, (4) SHAP provides business-interpretable explanations.

Empirical Findings: Risk discussion dominates healthcare M&A communications due to regulatory complexity. Communication strategies evolve significantly with market conditions—2021 featured technology-focused optimism, while 2024 emphasized execution capabilities. Integration focus emergence suggests market maturation.

Limitations: 193 transactions over four years limits generalizability. Analysis focuses only on formal announcements, missing broader stakeholder communications.

Future Research: Expand to additional sectors and time periods, incorporate post-announcement performance metrics, include regulatory filings and stakeholder communications for comprehensive analysis.

The methodology establishes a replicable framework for business communication analysis extending beyond sentiment analysis to capture domain-specific concepts.

Acknowledgments

I acknowledge the use of AI assistance (Gemini 2.5 Pro and Claude Sonnet 4) for code debugging and optimization suggestions during the development of this project. All core algorithmic implementations and analysis were developed independently. The AI assistance was primarily used for technical implementation support and does not substitute for the original research design, data collection, or critical analysis conducted by the author.

References

[1] Deloitte Life sciences and health care MA (2021-2025). *Life sciences and health care trends.*

[2] Lopez-Lira, A., Kwon, J., Yoon, S., Sohn, J.Y. & Choi, C. (2025). *Bridging language models and financial analysis.* arXiv preprint arXiv:2503.22693.

[3] Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). *Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV).* International Conference on Machine Learning, 2668-2677.

[4] Freeman, R. E. (1984). *Strategic management: A stakeholder approach.* Boston: Pitman Publishing.

[5] Noor, M.F., Juniar, A., Zulelli, R. & Hadi, A. (2025). *Critical Success Factors in Mergers and Acquisitions: Insights from a Systematic Review across Regions and Sectors.* Asian Journal of Management, Entrepreneurship and Social Science, 5(02), pp.699-720.